A ROLE FOR SETMAR IN GENE REGULATION:

INSIGHTS FROM STRUCTURAL ANALYSIS OF

THE DNA-BINDING DOMAIN IN COMPLEX WITH DNA

Qiujia Chen

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biochemistry and Molecular Biology
Indiana University

August 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Millie M. Georgiadis, Ph. D., Chair

_____
Thomas D. Hurley, Ph. D.

Doctoral Committee

_____
Ronald C. Wek, Ph. D.

June 30, 2016

_____
John J. Turchi, Ph. D.

_____
Mark R. Kelley, Ph. D.

DEDICATION

This thesis is dedicated to my parents Chuhui and Xuanna. Thank you for giving me the best education you could and great support. To my brother Jiamin, sister-in-law Xiaoqiong, and my niece Yuhan, for helping me survive all the stress.

I would like to dedicate this thesis to Dr. Jinsong Liu, who opened the door for me to the intriguing world of X-ray crystallography and structural biology.

Special dedication of this thesis must be given to Dr. Millie M. Georgiadis, a very talented and inspiring scientific mentor and advisor, for her patience, enthusiasm, and encouragement throughout this project.

ACKNOWLEDGEMENTS

Firstly, I would like to express deepest appreciation to my advisor, Professor Millie M. Georgiadis, for her continuous support of my Ph.D research work, for her motivation, patience, and immense knowledge. Her excitement in regard to protein-DNA crystallography helped me in all the time of research and writing of this thesis. Without her scientific guidance and persistent help, this thesis would not have been possible.

Besides my advisor, I would like to thank the rest of my thesis committee: Professors Thomas D. Hurley, Ronald C. Wek, John J. Turchi, and Mark R. Kelley, for their time, insightful comments, and encouragement, but also for the critical question which challenged me to widen my research from various perspectives.

In addition, a thank you to Dr. Suk-hee Lee, Dr. Quyen Hoang, Dr. Lan Chen, and Dr. Ronald C. Wek, who gave access to the laboratory and research facilities. Without their kind support it would not be possible to conduct this research.

I am thankful to my lab colleagues Dr. Hongzhen He, Isha Singh, and Sarah Delaplane, for all the fun we have had. I appreciate the help of my colleagues, Dr. Hyun Suk Kim, Michael Fusakio, Dr. Bibek Parajuli, Dr. Cindy Morgan, Dr. Tsuyoshi Imasaki, Dr. Kentaro Yamada, Krishna Kishore Mahalingan, Cameron Buchman, Dr. Wei Wang, and Dr. Yangshin Park. Also I thank my friends at Indiana University School of Medicine: Dr Jingling Liao, Dr. Jie Lan, Chen Chen, Liang Wang, and Chunxiang Wu, for their encouragement and moral support which make my stay and studies in Indianapolis more enjoyable.

I would like to extend my thanks to Dr. Mark Goebl and the Biochemistry office staff for their timely help and assistance. I would like to thank the beamline scientists at the GM/CA and SBC stations at the Advanced Photon Source, for their patience and scientific assistance.

I would also like to thank my family: my parents and my brother for supporting me spiritually throughout writing this thesis and my life in general.

Qiujia Chen

A ROLE FOR SETMAR IN GENE REGULATION: INSIGHTS FROM STRUCTURAL ANALYSIS OF THE DNA-BINDING DOMAIN IN COMPLEX WITH DNA

SETMAR is a chimeric protein that originates from the fusion of a SET domain to the *mariner Hsmar1* transposase. This fusion event occurred approximately 50 million years ago, after the split of an anthropoid primate ancestor from the prosimians. Thus, SETMAR is only expressed in anthropoid primates, such as humans, apes, and New World monkeys. Evolutionary sequence analyses have revealed that the DNA-binding domain, one of the two functional domains in the *Hsmar1* transposase, has been subjected to a strong purifying selection. Consistent with these analyses, SETMAR retains robust binding specificity to its ancestral terminal inverted repeat (TIR) DNA. In the human genome, this TIR sequence is dispersed in over 1500 perfect or nearly perfect sites. Given that many DNA-binding domains of transcriptional regulators are derived from transposases, we hypothesized that SETMAR may play a role in gene regulation. In this thesis, we determined the crystal structures of the DNA-binding domain bound to both its ancestral TIR DNA and a variant TIR DNA sequence at 2.37 and 3.07 Å, respectively. Overall, the DNA-binding domain contains two helix-turn-helix (HTH) motifs linked by two AT-hook motifs and dimerizes through its HTH1 motif. In both complexes, minor groove interactions with the AT-hook motifs are similar, and major groove

interactions with HTH1 involve a single residue. However, four residues from HTH2 participate in nucleobase-specific interactions with the TIR and only two with the variant DNA sequence. Despite these differences in nucleobase-specific interactions, the DNA-binding affinities of SETMAR to TIR or variant TIR differ by less than two-fold. From cell-based studies, we found that SETMAR represses firefly luciferase gene expression while the DNA-binding deficient mutant does not. A chromatin immunoprecipitation assay further confirms that SETMAR binds the TIR sequence in cells. Collectively, our studies suggest that SETMAR functions in gene regulation.

Millie M. Georgiadis, Ph. D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Bis-Tris	Bis (2-hydroxyethyl) iminotris (hydroxymethyl)-methane
ChIP assay	Chromatin immunoprecipitation assay
DBD	DNA-binding domain
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylene diamine tetra aceticacid
EMSA	Electrophoretic mobility shift assay
FA assay	Fluorescence anisotropy assay
FOM	Figure of merit
HTH	Helix-turn-helix
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LB	Luria-Bertani broth
MBS	Mariner binding site
My	Million years
Mya	Million years ago
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEG	Poly ethylene glycol
PMSF	Phenyl methyl sulfonyl fluoride

SAD	Single-wavelength anomalous dispersion
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SeMet	Selenomethionine
TCEP	tris(2-carboxyethyl)phosphine
TIR	Terminal inverted repeat
Tris	Tris (hydroxymethyl) aminomethane

INTRODUCTION

A. Human DNA transposons

Transposons, also called “jumping genes” or transposable elements, are mobile DNA sequences that are able to move themselves within the host genome (Wicker et al. 2007). Based on their transposition mechanisms, transposons can be classified into two

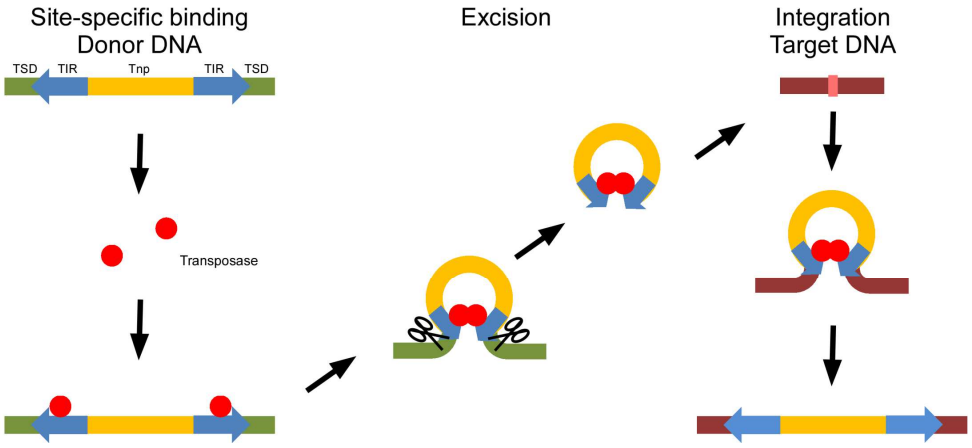

Figure 1. Cut-and-paste mechanism of DNA transposon.
A transposon contains a coding region for the transposase (Tnp, yellow rectangle), flanked on both ends by terminal inverted repeats (TIRs, blue arrows). The TIRs are flanked by target site duplications (TSDs), characteristic to each transposon family. The transposase protein (red sphere) specifically binds to its recognition TIRs at each end of the transposon. The transposase excises the transposon by cleaving the DNA at the ends of the TIRs following formation of a synaptic complex. The transposase recognizes a target site and integrates the transposon into the target DNA, upon which the target site gets duplicated. (Figure is adapted from Sinzelle, 2009)

classes: retrotransposons and DNA transposons. Unlike retrotransposons, which move via an RNA intermediate using a “copy-and-paste” mechanism, DNA transposons move-directly as a DNA intermediate by a “cut-and-paste” mode during transposition (Wicker et al. 2007, Sinzelle et al. 2009). Eukaryotic DNA transposons encode a single enzyme

called transposases to carry out the “cut-and-paste” transposition process (Lampe et al. 1996, Sinzelle et al. 2009) (Figure 1). Two functional domains are common in the DNA transposases: an N-terminal DNA-binding domain (DBD) that recognizes and binds to the terminal inverted repeats (TIR) on the transposon ends in a sequence-specific manner and a C-terminal catalytic domain that catalyzes both the DNA cleavage and strand transfer steps during transposition (Sinzelle et al. 2009). In the human genome, more than 380,000 DNA transposon copies have been identified and belong to 125 different families, accounting for approximately 3% of the whole genomic DNA (Lander 2001, Jurka et al. 2005, Bao et al. 2015). These human DNA transposons represent seven out

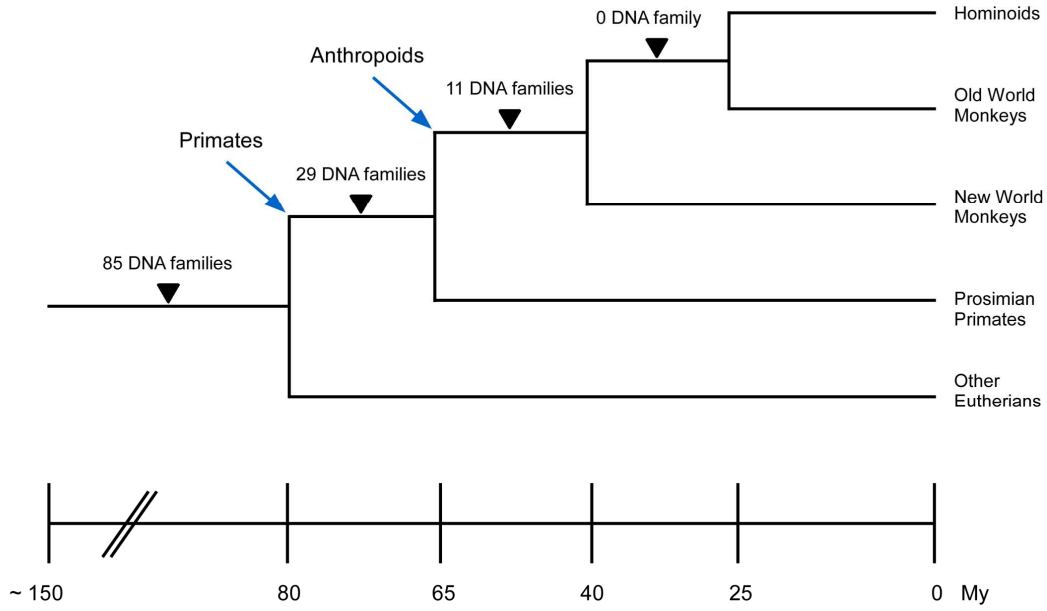

Figure 2. Summary of the activity of human DNA transposons through primate evolution. Solid triangles indicate different evolutionary time points when DNA transposable families were inserted in the human genome. 125 known human DNA transposable element families can be classified into three categories (from left to right): eutherian-wide (80-150 My), primate specific (65-80 My), and anthropoid specific (40-65 My). No DNA transposable elements were found to be active after the emergence of New World Monkeys. My: million years. (Figure is adapted from Pace and Feschotte, 2007)

of ten known superfamilies of eukaryotic DNA transposons (Jurka et al. 2005).

The evolutionary history of the human DNA transposons over the course of mammalian and primate evolution was analyzed comprehensively by Pace and Feschotte, using three independent computational methods to determine the average age of all 125 DNA transposon families identified in the human genome (Pace and Feschotte 2007). According to their analysis, eighty-five families were traced back to 80 million years ago (Mya), before the split of a primate ancestor from other eutherians (i.e. placental mammals). The remaining 40 DNA transposon families are primate-specific families and account for approximately 38 Mb of DNA in the human genome. Early primate evolution (65-80 Mya), that is after the divergence of a primate ancestor from the closest non-primate eutherian clades (mouse, rat, and rabbit) but prior to the emergence of prosimian primates (about 63 Mya), was a period of intense activity for DNA transposons, with the result that about 74,000 DNA transposons were inserted and fixed in the human genome (about 33 Mb of DNA). During the next phase of the primate radiation (40-65 Mya), i.e. after the split of prosimians, but prior to the emergence of New World monkeys, approximately 2,300 human DNA transposons were integrated during this period. However, there were no DNA transposon families significantly younger than the divergence of New World monkeys, approximately 40 Mya (Figure 2). Thus, the last active DNA transposon families have long been extinct and no functional DNA transposons are found in the human genome (Pace and Feschotte 2007).

B. *Hsmar1* discovery history and resurrection

As transposable elements can jump from one place to another in the genome, transposon insertion is the main cause of mutations in many animals. Insertion of Tc1 (*T*ransposon *C*aenorhabditis *elegans* number *1*) causes gene inactivation in several strains of *C. elegans* (Eide and Anderson 1985). Since the discovery of the Tc1 element, related elements discovered in other species were found to be homologous to Tc1. The best characterized example is the *mariner* element, first identified in *Drosophila mauritiana* (Jacobson et al. 1986, Jacobson 1990). Members of the Tc1/*mariner* superfamily of DNA transposons are found in a variety of organisms, ranging from fungi to humans (Robertson 1993, Robertson 1995, Robertson and Lampe 1995, Plasterk 1996, Robertson and Martos 1997, Robertson and Zumpano 1997, Lampe et al. 1999).

Human Tc1/*mariner* elements account for approximately one-third of all human DNA transposon copies (Smit and Riggs 1996, Smit 1999, Lander 2001). About two decades ago, researchers from several groups independently identified two quite different families of *mariner* elements in the human genome (Morgan 1995, Oosumi et al. 1995, Smit and Riggs 1996, Robertson and Martos 1997, Robertson and Zumpano 1997): *Hsmar1* (*H*omo *s*apiens *mar*iner *1*) (Robertson and Zumpano 1997) and *Hsmar2* (*H*omo *s*apiens *mar*iner *2*) (Robertson and Martos 1997). As the first representative of *mariner* elements recognized in mammalian genomes, the consensus sequence of *Hsmar1* was constructed by Robertson and colleagues (Robertson and Zumpano 1997). The consensus sequence of *Hsmar1* is 1287 bp long with 30 bp perfect TIRs and encodes

a 343 amino acid *mariner* transposase (Figure 3). As found in the Tc1/*mariner* superfamily, *Hsmar1* encodes only a single protein, the transposase, and is flanked at either end by the 30 bp TIRs. The putative *Hsmar1* transposase has two functional domains: a DNA binding domain and a catalytic domain. The DNA binding domain has two helix-turn-helix (HTH) motifs for TIR sequence-specific DNA binding. The catalytic domain contains the DD34D motif that is crucial for transposition and is another shared feature of eukaryotic *mariner* elements (Lampe et al. 1996, Lohe et al. 1997, Nowotny 2009, Montano and Rice 2011). The DD34D motif has first two aspartic acid residues (DD) separated from one another by about 90 residues followed by a third aspartic acid residue (D) at a distance of 34 residues.

In the same seminal study, the evolutionary history of *Hsmar1* copies in the human

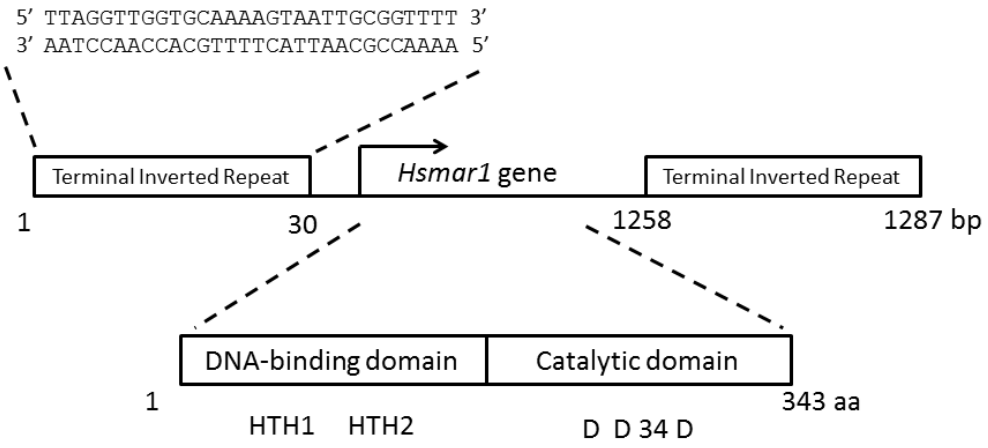

Figure 3. Schematic representation of *Hsmar1* transposon DNA and the encoded *Hsmar1* transposase protein.
The *Hsmar1* transposase gene is flanked by two 30-pb terminal inverted repeats (TIRs). The *Hsmar1* transposase contains a DNA-binding domain with two helix-turn-helix (HTH) motifs and a catalytic domain with a DD34D motif.

genome was also analyzed in detail by Robertson and colleagues (Robertson and Zumpano 1997). Ancestral *Hsmar1* was active during early primate evolution period (40-65 Mya) (Figure 2) and invaded early primate genomes approximately 50 Mya. Movement of the *Hsmar1* transposon was ongoing until at least 37 Mya, consistent with the evolutionary history of human DNA transposons (Pace and Feschotte 2007) (Figure 2). As a result of transposition, at least 200 copies of *Hsmar1* were integrated in the ancestral primate genome. Most of the current *Hsmar1* copies in the human genome have diverged from the consensus largely (by an average of 7.8% in DNA sequence) and have now lost transposition activity. However, one exceptional *Hsmar1* copy has been remarkably conserved, being only 2.4% divergent from the consensus *Hsmar1* gene. Interestingly, this *Hsmar1* copy retains the entire transposase coding region and is fused in-frame downstream of the *SET* (*Suppressor of variegation 3-9*, *Enhancer of zeste*, and *Trithorax* proteins of *Drosophila melanogaster*) gene coding region, forming a chimeric gene called *SETMAR*, which is the main topic of this thesis.

At the time the consensus sequence of *Hsmar1* was constructed, it was believed that a consensus sequence represents an active transposon sequence. When present in cells, this artificial DNA transposon is able to perform transposition. Based on this hypothesis, Ivics and colleagues successfully reconstructed two functional transposable elements: Sleeping Beauty from fish (Ivics et al. 1997) and Frog Prince from amphibians (Miskey et al. 2003). However, the consensus *Hsmar1* gene has no transposition function, assessed by the same transposition system established for Sleeping Beauty and Frog

Prince. After phylogenetic analysis, four amino acid substitutions were introduced in the consensus transposase protein sequence: C53R, P167S, L201V, and A219C. The resulting *Hsmar1* transposase protein functions as an active transposable element (Miskey et al. 2007).

C. The origin of *SETMAR* and its biological functions

As Robertson and colleagues revealed, ancestral *Hsmar1* was active during the early primate evolution period (40-65 Mya, Figure 2) and integrated at least 200 copies of *Hsmar1* in the ancestral primate genome (Robertson and Zumpano 1997). One particular copy of *Hsmar1* was fused in-frame to a preexisting *SET* gene forming a new primate chimeric gene *SETMAR* (Robertson and Zumpano 1997, Cordaux et al. 2006).Using sequence analysis and phylogenetic comparison, Cordaux and colleagues presented a proposed "molecular domestication" process of the *Hsmar1* transposase gene (Cordaux et al. 2006) (Figure 4). Approximately 50 million years ago, an ancestral *Hsmar1* transposon integrated downstream of the *SET* gene. Later, a retrotransposon *AluSx* inserted in the 5' TIR of the *Hsmar1* element, resulting in the deletion of 12 bp of the TIR and 4 bp of flanking genomic DNA. At a period of intense activity for *Hsmar1* transposon, the *AluSx* may have prevented this *Hsmar1* copy from mobilizing again since both TIRs of transposons are necessary for transposition. The next step towards the birth of *SETMAR* involved the capture and in-frame fusion of the *Hsmar1* transposase to the *SET* gene. Compared to tarsier, which only has the *SET* gene, all examined

anthropoid lineage *SETMAR* sequences share a 27-bp genomic deletion that removed the stop codon of the *SET* gene located at the 3' end of the second *SET* exon, creating a new intron as the current second intron of *SETMAR* (Figure 4). This process included a *de*

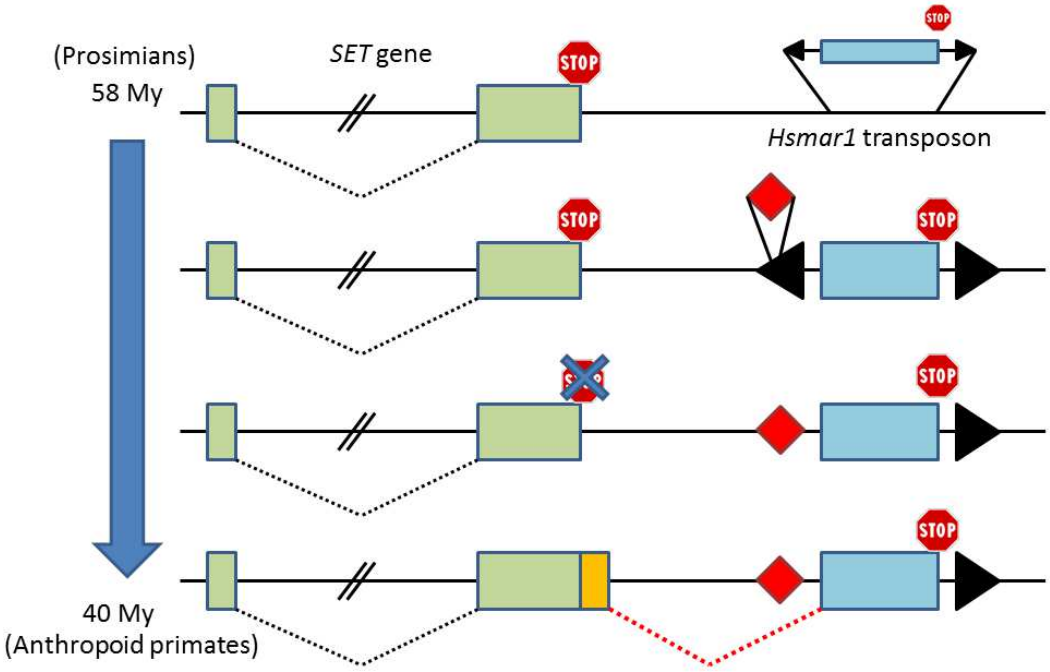

Figure 4. The Birth of *SETMAR*.

Series of proposed molecular events occurred within an evolutionary time window of less than 18 million years, after the emergence of the prosimians lineage and before the split of anthropoid primates (40-58 million years ago, Mya). A "*SET*-only" gene, which contains two *SET* exons, is conserved in all non-anthropoid species examined and terminated with a stop codon (stop sign). Two *SET* exons (light green boxes) are separated by a single intron (interrupted black line). An ancestral *Hsmar1* transposon entered downstream of the *SET* gene in the primate lineage around 50 Mya. The transposon shown here contains its TIRs (black triangles) and transposase coding sequence (blue box) with stop codon (stop sign). The secondary *AluSx* retrotransposon (red diamond) inserted in the 5' TIR of the *Hsmar1* element, immobilizing this *Hsmar1* copy. The deletion removed the ancestral stop codon of the *SET* gene (cross sign) and allowed the creation of the current second intron of *SETMAR*, which is represented as a dashed red line. This process was made possible by the *de novo* conversion from noncoding sequence to exonic sequence (yellow box). The *SETMAR* transcript now consists of three exons. SET: *Suppressor of variegation 3-9*, *Enhancer of zeste*, and *trithorax* proteins of *Drosophila melanogaster*. (Figure is adapted from Cordaux, 2006)

novo conversion of a 77-bp-long previously noncoding sequence into an exonic sequence, coding a linker between the end of the former SET protein and the beginning of the recruited *Hsmar1* transposase. Since this process took place after the divergence of an anthropoid ancestor from the prosimian primate clades, the *SETMAR* gene is only present in anthropoid primate lineages (humans, apes, Old World Monkeys, and New World Monkeys), but not in prosimians (Tarsier and Galago), and non-primate mammals (mouse, rat, dog, and cow).

In the human genome, the *SETMAR* gene is mapped on the short arm of chromosome 3p26.1, a region which is linked to a number of diseases, such as non-Hodgkin's lymphoma, acute leukemia, hereditary prostate cancer, myeloma, and myelodysplastic syndromes (Higgins et al. 2004). The SETMAR transcript, which consists of three exons, encodes a protein of 671 amino acids (Figure 5) (Note that the NCBI

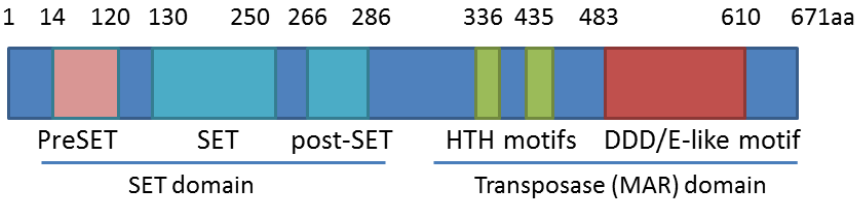

Figure 5. Schematic diagram of SETMAR protein.
The SET domain has three motifs: PreSET, SET, and post-SET. The transposase (MAR) domain is derived from the *Hsmar1* transposase, a member of the *mariner* superfamily of DNA transposon. The DDD/E-like motif in SETMAR is DD34N, with the last amino acid substituted from Asp or Glu to Asn. (Note that the NCBI reference sequence NP_006506 has been updated to NP_006506.3 with 13 amino acids extending at the N-terminal of SETMAR to total of 684 amino acids.)

reference sequence NP_006506 has been updated to NP_006506.3 with an additional 13 amino acids at the N-terminal of SETMAR resulting in a total of 684 amino acids. Numbering of amino acids is based on the 671 amino acids version in this thesis). Due to accumulating mutations within the transposase domain (Figure 6), SETMAR has lost its ability to function as a transposase (Liu et al. 2007, Miskey et al. 2007). However, SETMAR is a multi-functional protein which has lysine methylation activity, non-homologous end joining DNA repair activity, restart of stalled replication forks, and

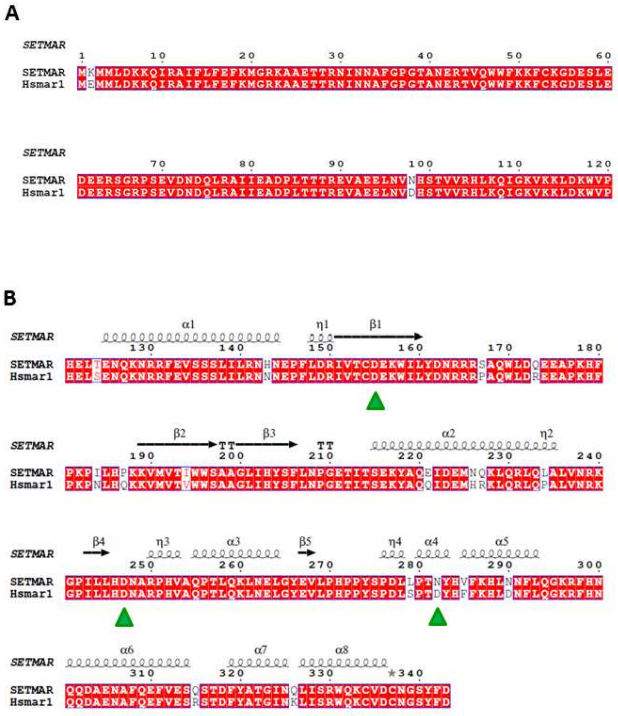

Figure 6. Protein sequence alignment of SETMAR transposase domain and predicted *Hsmar1* transposase.

A. The DNA-binding domain (DBD) of SETMAR has two mutations. B. The catalytic domain of SETMAR bares 17 mutations, in which the last amino acid of the DD34D motif in *Hsmar1*, Asp, is substituted to Asn. Green triangles represent DD34D motif in the catalytic domain of the *mariner* element.

chromosomal decatenation (Lee et al. 2005, Wray et al. 2009, Wray et al. 2009, De Haro et al. 2010, Wray et al. 2010, Carlson et al. 2015). SETMAR is overexpressed in the peripheral blood and bone marrow of acute myeloid leukemia (AML) patients as compared to that of healthy individuals (Jeyaratnam et al. 2014).

Evolutionary sequence analysis of SETMAR demonstrated that a continuous purifying selection has acted to preserve the DBD but not the catalytic domain of SETMAR (Cordaux et al. 2006). Protein sequence alignment of the SETMAR transposase domain and the consensus *Hsmar1* sequence also demonstrate that only two amino acid residues are substituted in the DBD. However, in the catalytic domain, 17 amino acids are different from the consensus ancestral sequence (Robertson and Zumpano 1997, Cordaux et al. 2006, Miskey et al. 2007) (Figure 6). Importantly, this domain now has a DD34N rather than the DD34D motif found in active *mariner* transposase and thought to play an important role in metal binding (Richardson et al. 2009, Goodwin et al. 2010, Kim et al. 2014). Consistent with this analysis, biochemical experiments showed that SETMAR has retained binding specificity to the putative *Hsmar1* TIR DNA, but has lost some of its catalytic abilities, resulting in an inactive DNA transposon in humans (Cordaux et al. 2006, Liu et al. 2007, Roman et al. 2007).

D. Crystal structures of DNA-binding domains of Tc1/*mariner* superfamily transposons

Members of Tc1/*mariner* superfamily share many features in terms of transposition mechanism. However, the TIRs differ in length and sequence, and the amino acid sequences of DBDs vary greatly for different members. For example, the Mos1 transposon (from *D. mauritiana*) contains a 28-bp TIR while Tc3 transposon (from *C. elegans*) has a 462-bp TIR at both ends of the cognate element (Watkins et al. 2004, Richardson et al. 2009) (Figure 7). The *Hsmar1*-TIR is a 30-bp DNA sequence (Robertson and Zumpano 1997) (Figure 3), within which a core 19-bp region was identified as a SETMAR-binding site by EMSA (Electrophoretic Mobility Shift Assay) (Cordaux et al. 2006) (Figure 8). Depending on the human genome sequence databases used, the number of

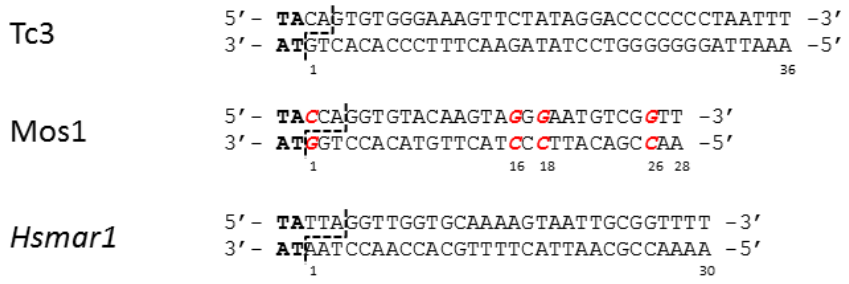

Figure 7. The sequences of the left end TIRs of Tc3, Mos1 and *Hsmar1* DNA transposons.
Tc3 has two binding sites separated by about 180 bp at each transposon end. Here the 36 bp outer left TIR of Tc3 transposon is shown. Mos1 contains 28 bp imperfect TIRs, differing from each other at positions 1, 16, 18, and 26, highlighted in red italics. *Hsmar1* has two identical TIRs on both ends. The TA target site duplication is shown in bold. The cleavage sites are marked by dotted lines. Note the differences of the cleavages sites.

potential SETMAR-binding sites ranges from 1,500 to 7,000 (Robertson and Zumpano 1997, Cordaux et al. 2006).

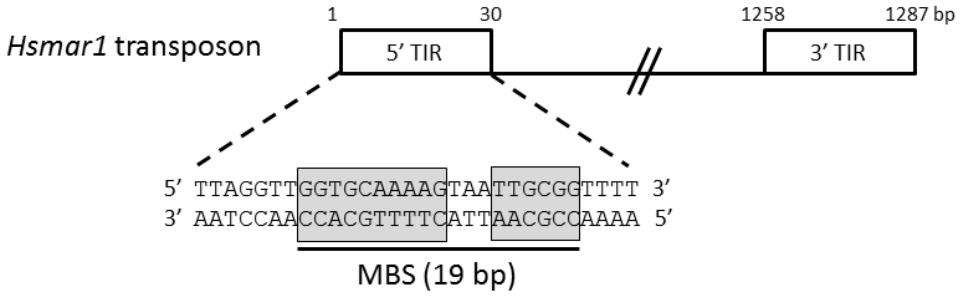

Figure 8. A 19-bp MAR (transposase domain) binding site (MBS) was identified by EMSA. A schematic representation of *Hsmar1* transposon is shown on the top, highlighting the 30-bp 5′ and 3′ TIR sequences. Black arrows show the directions of TIRs. Using purified recombinant protein, MBP-fused SETMAR transposase domain, and various TIR double-stranded oligonucleotides, Cordaux and colleagues identified a 19-bp MBS, which is shown as a shaded sequence. (Figure is adapted from Cordaux 2006)

Interestingly, the overall structural fold of the DBD is conserved among all members of Tc1/*mariner* superfamily, such as Mos1 and Tc3 (Watkins et al. 2004, Richardson et al. 2009) (Figure 9). Strikingly, the DBD of PAX6 (paired box protein 6), which shares a common ancestor with the current Tc1 family transposon (Breitling and Gerber 2000), shows a similar structural fold as well (Xu et al. 1999) (Figure 9). PAX6 is a well-characterized transcription factor that regulates the development of sensory organs and brain (Hanson et al. 1994, Azuma et al. 1996).

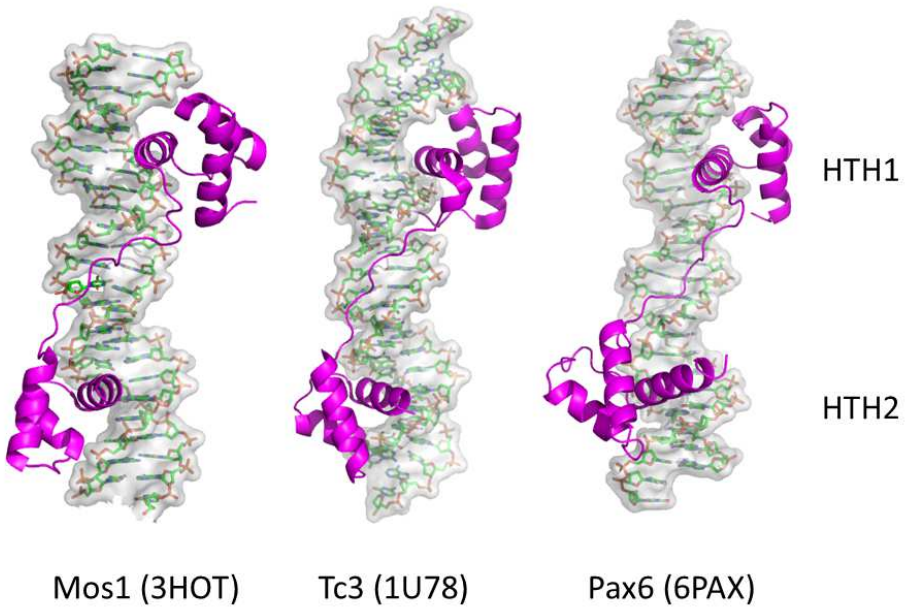

Figure 9. Crystal structures of DNA-binding domains of Tc1/*mariner* superfamily DNA transposons.
The DNA-binding domain contains two HTH motifs that bind to the major groove of DNA. A linker connecting two motifs interacts with the minor groove of DNA. PDB IDs are in parenthesis. Proteins are rendered as magenta ribbons. DNAs are represented as stick models with surface renderings (Xu et al. 1999, Watkins et al. 2004, Richardson et al. 2009).

E. Rationale and overview of this thesis

Although SETMAR retains robust sequence-specific DNA-binding activity, its biological significance is still unknown. The overall goal of this thesis is try to understand the structural basis of the DNA-binding activity and determine the associated biological function of SETMAR. Given that SETMAR retains robust DNA-binding activity and that the DNA-binding sites are dispersed throughout the human genome, SETMAR may use its ancestral transposase element, the DBD, to regulate expression of genes. To address

this possibility, I determined the crystal structures of the DBD complex with DNA and analyzed the protein-DNA interactions. The rationale for designing DBDs derived from SETMAR and different DNA sequences for crystallization is discussed along with phasing strategies, phasing experiments, model building, crystallographic refinement, and analysis of the structures. Using biochemical assays, I characterized the DNA-binding affinity and specificity of SETMAR based on the structural analysis. In these experiments, the relative contributions of nucleobase-specific interactions were assessed through amino acid substitutions in the DBD and base substitutions within the TIR or variant TIR sequences that were crystallized. To test the biological function of SETMAR, I conducted reporter gene regulation assays and ChIP assays with controls based on the structural analysis and DNA-binding results. Collectively, this work supports a role for SETMAR in transcriptional regulation.

MATERIALS AND METHODS

A. Protein expression and purification

For crystallographic studies, two different SETMAR DNA binding domains (DBDs) were used, one including residues 329-440, the other 316-440. To construct the first, DNA encoding amino acid residues 329-671 was PCR amplified from an existing pET15b vector containing the DNA for the SETMAR transposase domain (Goodwin et al. 2010) and subcloned into the pET28-derived pSUMO vector (Mossessova and Lima 2000) between restriction sites *Bam*H I and *Xho* I (Primers 329-671_F and 329-671_R). For the second plasmid construct, DNA encoding SETMAR residues 316-671 was PCR amplified from the pFLAG-CMV4 full-length SETMAR (wt) (Kim et al. 2014) and subcloned into the pSUMO vector between *Bam*H I and *Xho* I restriction sites (Primers 316-671_F and 316-671_R). Since the PCR amplification region encoding the DBD is short, about 300 bp in length, to ease the PCR product purification by gel extraction, the forward primer was designed to include the N-terminal 6xHis-SUMO tag coding region in the pET28-derived pSUMO vector. Using primers DBD_F and DBD_R, DNA encoding 6xHis-SUMO-SETMAR (aa 329-440) or 6xHis-SUMO-SETMAR (aa 316-440) was PCR amplified from cognate plasmid encoding SETMAR residues 329-671 or 316-671, respectively and subcloned into pSUMO vector between *Nco* I and *Xho* I restriction sites. Primers used for subcloning are listed in Table 1. Primers In this thesis, all primers were purchased from Integrated DNA Technologies, Inc., (Coralville, IA). The final plasmids were verified by DNA sequencing (Genewiz, Inc., South Plainfield, NJ). To optimize crystallization conditions,

mutations C381R and C381S were generated individually using the QuikChange II site-directed mutagenesis kit (Agilent Technologies, Santa Clara, CA). All the primers used to create the missense mutations are listed in Table 2.

Name	Sequence (5'—3')
329-671_F	ATATCGGATCC ATGAAAAT GATGTTAGACAAAAAGCAAATTCG
329-671_R	TATAGCTCGAGTTAATCAAAATAGGAACCATTACAATC
316-671_F	TATATCGGATCCGTGTTCCCCTCCTGCAAGCGATTGA
316-671_R	TATAGCTCGAGTTAATCAAAATAGGAACCATTACAATC
DBD_F	GATATACCATGGGCAGCAGCCATCATCATCATCATC
DBD_R	TATAGCTCGAGTTACACCTTTCCAATTTGCTTCAAATGTC

Table 1. Primers used for subcloning in the construction of SETMAR expression plasmids. Underlined letters are restriction enzyme sites, engineered into the oligonucleotide primers.

For experimental selenomethionine (Se-Met) single anomalous dispersion (SAD) phasing purposes, several Met substitutions were introduced to ensure a robust Se anomalous signal. Mutations were made by using the QuikChange II site-directed mutagenesis kit (Agilent Technologies, Santa Clara, CA) in the 329-440(C381R) plasmid to make pET28b-SUMO-SETMAR-329-440(C381R) double mutant constructs (See Table 2 for primer information). Of these expressed SETMAR recombinant proteins, SETMAR (C381R)(I359M)(L423M) was stably expressed in sufficient quantities for crystallization experiments.

Name	Primer	Sequence (5′ – 3′)
C381R	Forward	GCAGTGGTGGTTCAAGAAGTTTCGCAAAGGAGATG
	Reverse	CATCTCCTTTGCGAAACTTCTTGAACCACCACTGC
C381S	Forward	GCAGTGGTGGTTCAAGAAGTTTAGCAAAGGAGATG
	Reverse	CATCTCCTTTGCTAAACTTCTTGAACCACCACTGC
L343M	Forward	AAGCAAATTCGAGCAATTTTCATGTTCGAGTTCAAAATGGGTCGT
	Reverse	ACGACCCATTTTGAACTCGAACATGAAAATTGCTCGAATTTGCTT
I359M	Forward	CAGAAACAACTCGCAACATGAACAATGCATTTGGCC
	Reverse	GGCCAAATGCATTGTTCATGTTGCGAGTTGTTTCTG
L404M	Forward	GAAGTTGACAACGACCAGATGAGAGCAATCATCGAAG
	Reverse	CTTCGATGATTGCTCTCATCTGGTCGTTGTCAACTTC
L423M	Forward	CACGAGAAGTTGCTGAAGAAATGAATGTCAACCATTCTACGGT
	Reverse	ACCGTAGAATGGTTGACATTCATTTCTTCAGCAACTTCTCGTG
R371A	Forward	GCCCAGGAACTGCTAACGAAGCTACAGTGCAGTGG
	Reverse	CCACTGCACTGTAGCTTCGTTAGCAGTTCCTGGGC
S428A	Forward	GAACTCAATGTCAACCATGCTACGGTCGTTCGACATT
	Reverse	AATGTCGAACGACCGTAGCATGGTTGACATTGAGTTC
R432A	Forward	ACCATTCTACGGTCGTTGCACATTTGAAGCAAATTGG
	Reverse	CCAATTTGCTTCAAATGTGCAACGACCGTAGAATGGT

Table 2. Primers used for site-directed mutagenesis (one letter amino acid codes are used for the wild-type and substituted residues). Forward and reverse primers for each site were used in the same PCR reaction. Underlined bases represent the mutated codon.

Se-Met protein was expressed using an adapted protocol from literature (Van Duyne et al. 1993). *Rosetta* cells (EMD Millipore, Billerica, MA) were grown in M9 minimal media. At OD 0.5-0.6, an amino acid cocktail (100 mg/L culture of Lys, Phe and Threonine, 50 mg/L culture of Ile, Leu and Val) solution was added to inhibit methionine synthesis. Selenomethionine was supplied to the media at a final concentration 60 mg/L. The SETMAR proteins were induced in the bacterial cells by addition of 1 mM IPTG and grown at 18 ℃ overnight. The Se-Met protein purification protocol was the same as the wild-type protein, described below.

The DBDs of SETMAR used in this study including residues 329-440 or 316-440, and various substituted versions of these constructs were expressed in *Escherichia coli* (*E. coli*) *Rosetta* cells and purified as previously described (Kim et al. 2014). In brief, following lysis and centrifugation, the cell lysate was applied to a Ni-NTA column (Qiagen, Valencia, CA) and then subjected to on-column cleavage with the SUMO-specific Ulp1 protease to remove the N-terminal His-SUMO affinity tag. The eluent was then applied to a tandem Q-Sepharose/SP-Sepharose column (GE Healthcare, Marlborough, MA). The protein was eluted from the SP-Sepharose column by using a NaCl gradient from 50 mM to 1 M, subjected to size exclusion gel chromatography, and then concentrated to approximately 5 mM by using 10 kDa molecular weight cutoff concentrators (EMD Millipore, Billerica, MA). The protein was stored in 50 mM HEPES pH 7.5, 500 mM NaCl, and 1 mM DTT at -80 ℃.

For biochemical studies, the full-length SETMAR(wt) gene was cloned into the *Nde* I/*Xho* I site of pET15b (EMD Millipore, Billerica, MA), which was constructed by a former post-doc in our lab, Dr. Kristie D. Goodwin. Into this plasmid, mutations resulting in R371A, S428A, and R432A variants of the full-length SETMAR were generated individually by using the QuikChange II site-directed mutagenesis kit (Agilent Technologies, Santa Clara, CA). The primers used for PCR amplification are listed in Table 1. The full length SETMAR(wt) and mutants were expressed in *Rosetta* cells (EMD Millipore, Billerica, MA) and induced by culturing at 20 ℃ overnight with 0.1 mM IPTG and 50 μM $ZnCl_2$ (Carlson et al. 2015). Cells were lysed in 50 mM phosphate, pH 7.8, 300 mM NaCl, and 10 mM imidazole by French press (Aminco), and the supernatant after ultracentrifugation at 35,000 rpm for 30 min at 4 ℃ was loaded onto a Ni-NTA resin column. Eluted protein from the Ni-NTA column was diluted to 60 mM NaCl with no salt buffer (50 mM Tris-Cl, pH 7.0, 1 mM DTT) and applied to a Q-Sepharose ion exchange column. Elution peak fractions were collected, concentrated, and applied to a gel filtration column (Superdex 200 16/60 prep, GE Healthcare, Marlborough, MA) buffered with 50 mM Tris-Cl pH 7.0, 500 mM NaCl, and 1 mM DTT. The full length SETMAR(wt) and substituted proteins were concentrated using 10 kDa molecular weight cutoff concentrators (EMD Millipore, Billerica, MA). Protein was stored in 50 mM Tris-Cl pH 7.0, 500 mM NaCl, 1 mM DTT buffer at -80 ℃.

B. DNA oligonucleotides for crystallization

The oligonucleotides used in this study were synthesized by Midland Certified Reagent Company, Inc. (Midland, TX) on a one micromole scale and purchased in gel-purified, desalted form, and used without further purification. In all, eight duplex oligonucleotides were screened in crystallization experiments (Figure 11 and Figure 13). For the TIR complex, two oligonucleotides, 5'– GGTTGGTGCAAAAGTAATTGCGGTTA –3' and its complementary strand 5'– AACCGCAATTACTTTTGCACCAACCT -3', were annealed to form a 25-mer duplex DNA, TIR2 in Figure 11, which featured overhanging 3' A and T, respectively. Similarly, variant-TIR1 DNA shown in Figure 13 was prepared by annealing complementary oligonucleotides 5'- GCAGTGTGCAAAAGTGATTGCGGCTA -3' and 5'-AGCCGCAATCACTTTTGCACACTGCT -3'. For experimental phasing, the underlined "Ts" were replaced by 5-BrdU.

C. Crystallization

All of the variations of the DBD (329-440 (wild-type), 329-440 (C381S), 329-440 (C381R), and 316-440 (C381S)), were mixed with duplex DNA (5 mM stock) to make a final protein:DNA molar ratio of 1:1.2 in 50 mM HEPES pH 7.5, 150 mM NaCl, and 1 mM DTT. The resulting protein concentration was 500 μM. The protein-DNA complex was incubated on ice for 15 min prior to crystallization.

Initial crystallization screens were performed using the Art Robbins Gryphon crystallization robot (Art Robbins Instruments, Sunnyvale, CA) with 0.6 μL drops (0.3 μL

complex plus 0.3 μL reservoir solution) and 60 μL reservoirs in 96 well sitting drop vapor diffusion plates (Intelli-Plate 96-3 LVR, HR3-185; Hampton Research, Aliso Viejo, CA). Subsequently, all crystals were grown by vapor diffusion in 2 μL (1 μL complex plus 1 μL reservoir solution) hanging drops at 20 ℃ suspended over 500 μL of reservoir solution. The crystals for data collection were obtained by micro-seeding, cryocooled in a solution containing 20% ethylene glycol added to a stabilizing solution, and flash frozen in liquid nitrogen before data collection.

D. Data collection and data processing

Diffraction data were collected at 100 K at the 23-ID-B, 23-ID-D, and 19-ID beamlines at the Advanced Photon Source, Argonne National Laboratory. For experimental phasing, single-wavelength SAD data sets were collected from BrdU-labeled or SeMet/BrdU labeled protein-DNA complex crystals at the bromine or selenium absorption peak wavelength, 0.91922 Å and 0.97938 Å, respectively. Diffraction data were processed using XDS (McCoy et al. 2007) at 23-ID beamlines or HKL3000 (Minor et al. 2006) at 19-ID. Statistics for data processing and crystallographic refinement statistics are summarized in Table 5.

E. Experimental phasing and structure determination

For phasing purposes, Se SAD data (TIR complex, Se SAD in Table 5) were collected to 2.66 Å for DBD 329-440 (C381R)(I359M)(L423M) complexed with BrdU substituted TIR

DNA. Using AutoSol (Adams et al. 2010) primed with the three sites identified in the initial phasing experiment described in the Results section, five Se sites were identified; phases calculated from these sites resulted in a very interpretable electron density map. Autobuild (Adams et al. 2010) functions within AutoSol (Adams et al. 2010) were used to obtain a partial model of the DNA and two HTH motifs. A model containing residues 330-437 and the complete DNA duplex was completed through iterative cycles of model building in COOT (Emsley et al. 2010) and crystallographic refinement in REFMAC (Murshudov et al. 1997, Winn et al. 2011) and PHENIX (Adams et al. 2010). Although Br SAD datasets collected for the DBD complexed with BrdU-containing DNA crystals did not prove useful for phasing purposes, they were useful in confirming the positions of the BrdU in the DNA model. The positions of the SeMet residues were confirmed by anomalous difference Fourier analysis.

Diffraction data for the TIR complex (TIR complex, High Resolution in Table 5), including DBD 329-440 (C381R) (I359M) (L423M) complexed with brominated TIR DNA were collected to 2.37 Å. The structure was determined by molecular replacement in PHASER (McCoy et al. 2007, Winn et al. 2011) using the above model derived from the Se SAD experimental electron density map as the search model. A final refined model was obtained following iterative cycles of model building in COOT (Emsley et al. 2010) and refinement in PHENIX (Adams et al. 2010) and BUSTER (Bricogne G. and Roversi P 2016) using individual atomic coordinates and B-factors, maximum likelihood targets, and TLS parameters. Based on analysis from the TLS Motion Determination server

(http://skuld.bmsc.washington.edu/~tlsmd/index.html)(Painter and Merritt 2006, Painter and Merritt 2006), both TIR and variant-TIR complexes can be partitioned into 6 TLS groups: chain A 331-396, chain A 397-437, chain B 1-15, chain B 16-26, chain C 1-10, and chain C 11-26.

The crystal structure of the DBD protein containing residues 316-440 (C381S) complexed with brominated variant-TIR DNA (Variant-TIR complex in Table 5) was determined at 3.07 Å by molecular replacement as implemented in PHASER (McCoy et al. 2007, Winn et al. 2011) using the refined TIR complex as the search model and refined similarly to the TIR complex using both REFMAC (Murshudov et al. 1997, Winn et al. 2011) and PHENIX (Adams et al. 2010).

F. Fluorescence anisotropy assays (FA assay)

FA assays were conducted as described previously (Kim et al. 2014). In brief, 20 nM rhodamine-labeled DNA probe was incubated with varying concentrations of protein in a 50 µl reaction mixture buffered in 50 mM HEPES, pH 7.0, 150 mM NaCl and 1 mM DTT. Oligonucleotides were ordered from Midland Certified Reagent Company, Inc. (Midland, TX). We measured fluorescence anisotropy data by using the Envision$^{[TM]}$ 2102 Multilabel Plate Reader (PerkinElmer Life Science, Waltham, MA) in the Chemical Genomics Core Facility of Indiana University School of Medicine. K_D values were calculated by fitting the data to a one-site binding saturation ligand binding curve (SigmaPlot, version 11.2). Every titration experiment was conducted three times independently with triplicate

readings each time. A 5'- (rhodamine)(C6 amino) – AACCGCAATTACTTTTGCACCAACCTAA -3' oligonucleotide was annealed to its complementary sequence to make the *Hsmar1* TIR duplex DNA probe. For the *Hsmar1* variant TIR duplex DNA probe, an oligonucleotide 5'– (rhodamine)(C6 amino) – AGCCGCAATCACTTTTGCACACTGCTAA –3' was annealed to its complementary sequence. The DNA sequence of the *Hsmar1* variant TIR probe is the same as the one in the variant-TIR complex.

G. Protein-DNA binding competition assay

Competition assays were performed by titrating a protein-rhodamine DNA solution with an unlabeled DNA duplexes obtained from Midland Certified Reagent Company, Inc. (Midland, TX). Unlabeled DNA duplexes were prepared and annealed as described above for the fluorescently labeled oligonucleotides. Duplexes that successfully competed with the labeled probe displaced the fluorescent probe resulting in a loss of fluorescence anisotropy. For this study, we used 300 nM FL SETMAR (wt) with 20 nM DNA probe. The concentrations of protein and DNA probe were determined from the binding assay in which 70% of the saturated fluorescence anisotropy was measured. The buffer used in this assay was the same as above: 50 mM HEPES, pH 7.0, 150 mM NaCl and 1 mM DTT. The data generated from this analysis were plotted as a function of anisotropy against log of the unlabeled DNA concentration. DNA sequences used in this competition assay are listed in Table 6.

H. Luciferase Reporter Assays

To construct a luciferase reporter plasmid, five tandem repeats of *Hsmar1* TIR sequence were cloned in upstream of an SV40 promoter of the pGL3-Promoter Vector (Promega, Madison, WI; cat. E1761) at an *Xho* I site, generating pGL3-promoter-5xTIR. The tandem repeat DNA sequence was purchased from IDT gene synthesis, shown in Figure 10. The expression plasmid, pFLAG-CMV4-SETMAR (wt), contains full-length SETMAR (aa 1–671) with a FLAG epitope tag at the N terminus (Sigma-Aldrich, St. Louis, MO). Using the QuikChange II site-directed mutagenesis kit (Agilent Technologies, Santa Clara, CA), pFLAG-CMV4-SETMAR (R371A) was constructed using the above wild-type plasmid as a template.

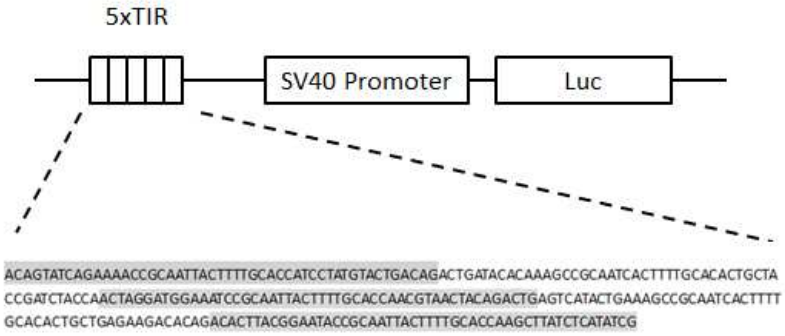

Figure 10. Schematic diagram of five tandem repeats of *Hsmar1* TIRs in the pGL3-promoter luciferase vector.

In collaboration with Mike Fusakio in Dr. Ron Wek's lab, HEK293T cells were grown in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% FBS, and Penicillin/Streptomycin (5ml in 500mL) (HyClone™, GE Healthcare, Marlborough, MA;

cat. SV30010) at 37℃ in a 5% CO_2 incubator. Cells were transfected using FuGENE 6 transfection reagent (Promega, Madison, WI). For one well of a 6- well cell culture plate, each transfection mixture contained 1 μg of luciferase reporter plasmid, 1 μg of expression plasmid, and 0.05 μg of pRL-CMV Renilla luciferase plasmid as an internal control. After 48 h, the transfected cells were lysed and assayed using the Dual-luciferase reporter assay system (Promega, Madison, WI). The firefly and Renilla luciferase activities were measured by 20/20^n Luminometer (Turner BioSystems, Sunnyvale, CA). Relative luciferase activity was calculated as the firefly luciferase divided by *Renilla* luciferase. The assays were performed three times with triplicate measurements for each. Statistical significance was determined by Student's t-test in GraphPad (Prism 6).

I. Chromatin immunoprecipitation (ChIP) assay

For one 10-cm cell culture dish, HEK293T cells were transfected with 6 μg of expression plasmid pFLAG-CMV4-SETMAR (wt) using FuGENE 6 transfection reagent (Promega, Madison, WI). One immunoprecipitation required 10 to 15 million cells. After 48 h, the transfected cells were fixed in 1% formaldehyde by adding 650 μL of 16% formaldehyde (Thermo Scientific, Waltham, MA; cat. 28906) to each 10-cm culture dish containing 10 mL medium. The cross-linking of proteins to DNA was carried out at room temperature for 10 min and was stopped by adding 1 mL of 10X glycine to cells at a final concentration of 125 mM. After washing twice with PBS, two cell dishes (about 10-15

million cells) were scraped and combined into one 1.5 mL tube. Cells were resuspended in 1.2 mL lysis buffer containing 50 mM Tris-pH7.5, 150 mM NaCl, 5 mM EDTA, 1% TritonX-100, and protease inhibitors (1 mM PMSF, 1X protease inhibitor cocktail Set V (EMD Millipore, Billerica, MA; cat. 539127). Chromatin in the lysate was sonicated using a Bioruptor 300 (Diagenode, Denville, NJ) to an average length of 200-500 bp as determined by agarose gel electrophoresis. The sonication condition was 30 cycle number, 30 seconds time on/off at high power. Immunoprecipitation (IP) was performed using Anti-FLAG M2 affinity agarose (Sigma-Aldrich, St. Louis, MO; cat. A2220) for SETMAR and mouse IgG-agarose (Sigma-Aldrich, St. Lois, MO; cat. A0919) for a negative control group. Before IP, 40 µL (50% slurry) was washed with buffer containing 50 mM Tris-pH7.5, 150 mM NaCl, 5 mM EDTA and 1% TritonX-100. For IP, 500 µL of supernatant of the sheared chromatin was added to the washed agarose beads and rotated overnight at 4 ℃. For each sample, agarose beads were collected at 1,000 rpm and washed three times in low salt wash buffer (50 mM Tris-pH7.5, 150 mM NaCl, 5 mM EDTA, and 1% TritonX-100), one time in high salt wash buffer (50 mM Tris-pH7.5, 500 mM NaCl, 5 mM EDTA, and 1% TritonX-100), and then one time in TE buffer (pH 8.0). Following 5 min of mild shaking at 4 ℃, each wash was centrifuged at 1,000 rpm at 4 ℃, and the supernatants were discarded. Immunoprecipitated complexes were incubated with 150 µL of 0.1 M glycine (pH 3.5) for 5 min at room temperature. The resin was centrifuged for 30 seconds at 1,000 rpm at room temperature to elute complexes. The supernatant was transferred to a new tube containing 15 µL of 10X neutral buffer (0.5 M

Tris pH 7.5, 1.5 M NaCl). To reverse crosslinks and digest protein, 2 μL proteinase K (20 mg/mL, Ambion, Foster City, CA; cat. AM2546) was added and incubated at 65 ℃ overnight. DNA samples were purified by using a PCR purification kit (Qiagen, Valencia, CA) and eluted with 30 μL EB buffer in the final step.

To detect SETMAR enrichment in the *TOPBP1* promoter region, primer pairs, which amplify the *TOPBP1* promoter region (from -188 to -46 bp), were 5'- CGTTTGACATTTCCGCTCTTCTGCTGC -3' and 5'- CCTACCCCAAAGCAAACGCTGGAGAA -3'. PCR mixtures contained 5 μL of IP DNA, 2 μL of primer pairs (10 μM of each), 10 μL of 2X SYBR-Green Reaction Mix (Bioline USA Inc, Taunton, MA) and 3 μL ddH_2O in a total volume of 20 μL. qPCR was performed at 95 ℃ for 3 min, 40 cycles of denaturation (95 ℃ for 15 sec) and annealed/extended at 60 ℃ for 60 sec. The signal ratio was calculated using 2^(C[T] IgG - C[T] Anti_FLAG) and obtained from three independent ChIP experiments.

J. Plasmid DNA immunoprecipitation

The above ChIP protocol was modified for plasmid DNA immunoprecipitation. HEK293T cells were transfected pGL3-promoter-5xTIR and pFLAG-CMV4-SETMAR (wt). The cross-linking, sonication and immunoprecipitation steps are the same as above.

qPCR was used to detect binding of SETMAR to the plasmid DNA. PCR mixtures contained 5 μL of IP DNA, 2 μL of primer pairs (10 μM of each), 10 μL of 2X SYBR-Green Reaction Mix (Bioline USA Inc, Taunton, MA) and 3 μL ddH_2O in a total volume 20 μL.

Primer sequences 5’- GAGCTCTTACGCGTGCTAG -3’, and 5’- TAATTGAGATGCAGATCGCAGAT -3’, were designed to anneal to the multi-cloning sites of pGL3-promoter flanking the 5xTIR binding site (Promega, Madison, WI). qPCR was performed at 95 ℃ for 3 min, 40 cycles of denaturation (95 ℃ for 15 sec) and annealed/extended at 60 ℃ for 60 sec. The signal ratio was calculated using 2^(C[T] IgG - C[T] Anti_FLAG) and obtained from three independent experiments.

K. Western blot analysis

In collaboration with Mike Fusakio in Dr. Ron Wek’s lab, western blot analysis was conducted to evaluate the protein expression stability of mutant SETMAR(R371A) in HEK293T cells, as described previously (Teske et al. 2013). Cells were transfected with pFLAG-CMV4-SETMAR (wt) or pFLAG-CMV4-SETMAR (R371A). After 48 h, the transfected cells were lysed in a RIPA-buffered solution containing 50 mM Tris-HCl (pH 7.9), 150 mM sodium chloride, 1% Nonidet P-40, 0.1% SDS, 100 mM sodium fluoride, 17.5 mM β-glycerophosphate, 0.5% sodium deoxycholate, and 10% glycerol supplemented with EDTA-free protease inhibitor cocktail tablet (Roche, Indianapolis, IN). Protein was separated via SDS–polyacrylamide gels (SDS-PAGE). Primary antibodies used in this study included those against β-actin (A5441; Sigma-Aldrich, St. Louis, MO) and Anti-FLAG M2 (F1804; Sigma-Aldrich, St. Louis, MO) were prepared at a concentration of 1:500 in 5% milk powder in PBS. Secondary antibodies were purchased from Bio-Rad (Hercules, CA). FLAG protein was detected through chemiluminescence as previously

described (Teske et al. 2013). Actin protein was recorded using the Odyssey infrared imaging system (LI-COR Biosciences, Lincoln, NE).

RESULTS

A. Variation of the protein and DNA components of the complex

The ability to crystallize protein-DNA complexes involving a small DBD relative to the size of the DNA will necessarily depend on optimizing both the protein and DNA components of the complex. In this study, four different variants of the DBD and eight different oligonucleotides were screened in crystallization trials. Variants of the DBD include the following: wild-type 329-440, 329-440(C381S), 329-440(C381R), and 316-440(C381S). The initial DBD including the wild-type SETMAR sequence was defined

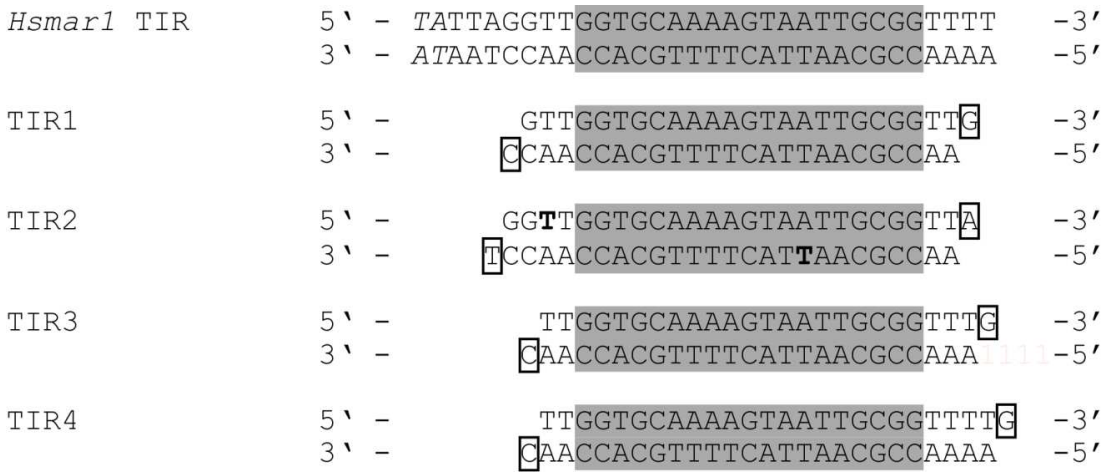

Figure 11. *Hsmar1* TIR based DNA sequences used for crystallization.
The left end *Hsmar1* TIR sequence is shown on the top. The italic TA is the characteristic target site duplication in the Tc1/*mariner* transposon superfamily. Four different TIR-derived sequences were used for initial crystallization trials. The core 19-bp binding site is in gray. In TIR2, the bold Ts represent positions for which T was replaced by 5-BrdU in the Br-DNA duplexes. The overhang nucleotides in boxes are designed for facilitating crystal packing.

as residues 329-440 based on a sequence comparison with the related transposase, Mos1, and was complexed to four different TIR-containing DNA duplexes for crystallization trials. The four oligonucleotides shown in Figure 11 were designed based on prior knowledge of the *Hsmar1* binding site within the TIR (Cordaux et al. 2006) and the available crystal structures Tc3 and Mos1 complexed with DNA (Watkins et al. 2004, Richardson et al. 2009). In these initial designs, the placement of the 19 bp mariner binding site (MBS) relative to the ends of the oligonucleotides and the length of the duplex regions were varied. The DNA sequences include two 24-mer duplexes with one 5' overhanging nucleotide on each strand, TIR1 and TIR3, and two 25-mer duplexes with one 5' overhanging nucleotide, TIR2 and TIR4. In TIR1, the 19 bp MBS was placed centrally with three additional 5' nucleotides from the TIR sequence and two nucleotides 3' of the recognition element. TIR3 included two 5' nucleotides and three 3' nucleotides on either side of the MBS. TIR2 included four 5' nucleotides and two 3' nucleotides, and TIR4 two 5' nucleotides and four 3' nucleotides.

To identify crystallization conditions, approximately 400 conditions contained within Index, Natrix and Natrix2, Crystal Screen and Crystal Screen 2, and PEG-ion and PEG-ion2 kits from Hampton Research, Inc. were screened. Crystals were obtained for the complex of wild-type DBD with TIR2 from the Index G2 condition (Hampton Research), which contains 0.2 M Li_2SO_4, 0.1 M Bis-Tris pH 5.5, and 25% w/v PEG 3350 (Figure 12 A). During optimization of this crystallization condition, we found that addition of TCEP, or tris(2-carboxyethyl)phosphine, improved the crystal size and appearance (Figure 12 B).

However, the crystals grown in the presence of TCEP only diffracted X-rays to about 8 Å. Since TCEP is highly effective at keeping sulfhydryls reduced, limiting disulfide bond formation, we then looked for cysteine residues in the protein sequence that might form intermolecular disulfide bonds during the crystallization process and identified one Cys residue, C381, within this DBD. The equivalent position in Mos1 is Lys54, which is solvent exposed in the crystal structure of the paired end Mos1 complex (Richardson et al. 2009). Interestingly, it was this same cysteine that was substituted with Arg, together with other critical substitutions, to reconstruct an active *Hsmar1* transposase from the

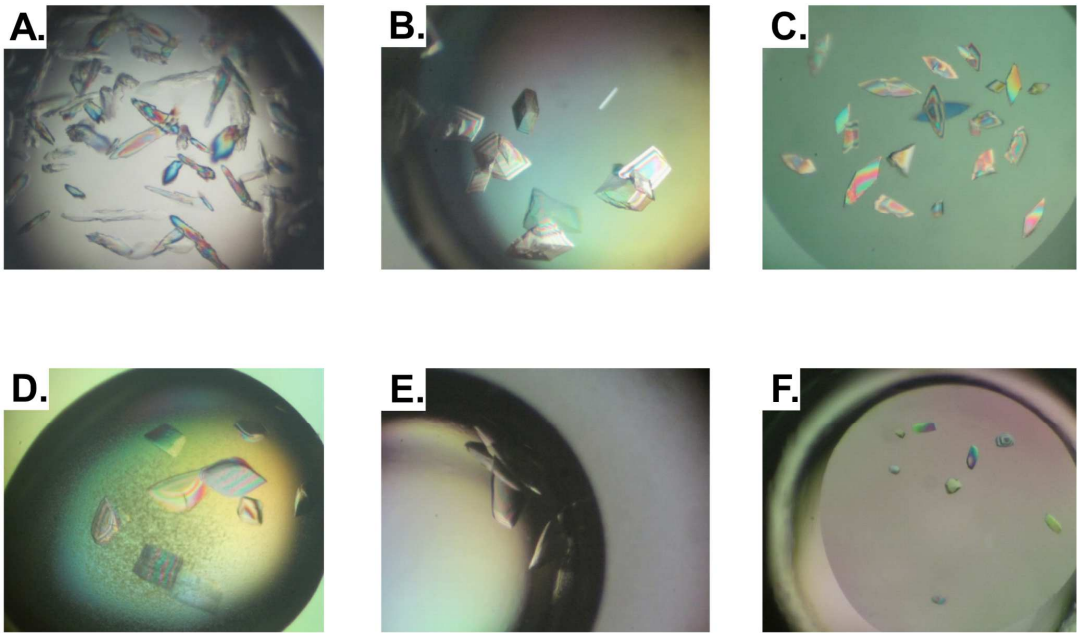

Figure 12. Crystal images of initial crystallization trials.
(A) 329-440(wt) with TIR2. (B) 329-440(wt) with TIR2 with TCEP additive. (C) 329-440(C381S) with TIR2. (D) 329-440(C381S) with variant TIR1. (E) 316-440(C381S) with variant TIR1. (F) 329-440(C381R) with TIR2.

consensus sequence (Miskey et al. 2007). Thus, we introduced two different substitutions for C381, C381S, representing a relatively conservative substitution, and C381R, as found in the revertant SETMAR, for further crystallization trials.

Crystals of DBD 329-440(C381S) complexed with TIR2 were similar in morphology to the wild-type DBD:TIR2 complex grown in the presence of TCEP (Figure 12 B and C). The crystallization condition Index E11 (Hampton Research) contained 25% poly (acrylic acid sodium) 5100, 0.10 M HEPES pH7.5, 0.02M $MgCl_2$. A very small cryocooled crystal diffracted X-rays to 4 Å and belonged to the monoclinic space group C2, with unit-cell parameters a=78.7, b=168.6, c=74.9 Å, β=108.35°. In this case, attempts to cryocool larger crystals were unsuccessful with diffraction limited to low resolution.

Another variable screened with the goal of improving the crystals was the sequence of the DNA used for crystallization. Changes in the nucleotide sequence of the oligonucleotide duplexes were introduced by attempting to retain critical nucleobases identified in the EMSA analysis (Cordaux et al. 2006) while varying others that might not be critical for SETMAR binding (Figure 13). Our screen of these variant-TIR oliognucleotides complexed with 329-440(C381S) produced crystals in the Natrix A2 condition (0.01 M Magnesium acetate, 0.05 M MES pH5.6, 2.5 M ammonium sulfate) (Figure 12 D). However, cryocooling of the crystals proved problematic as indicated by smeared diffraction patterns that extended only to low resolution.

Realizing that the transposase is fused to the C-terminus of a SET domain, it occurred to us that additional residues derived from the linking region between the SET and transposase domains might be involved in DNA-binding. Accordingly, a DBD including residues 316-440 with the C381S substitution was screened with both TIR and variant TIR-containing oligonucleotides. Crystals were obtained for a complex including 316-440(C381S) with variant TIR 1 in 0.10 M magnesium formate and 15% PEG3350 (Figure 12 E). Crystals for this complex diffracted to 3.15 Å and belonged to space group $C222_1$ with cell dimensions of a=72.23 Å, b=164.39 Å, c=67.96 Å.

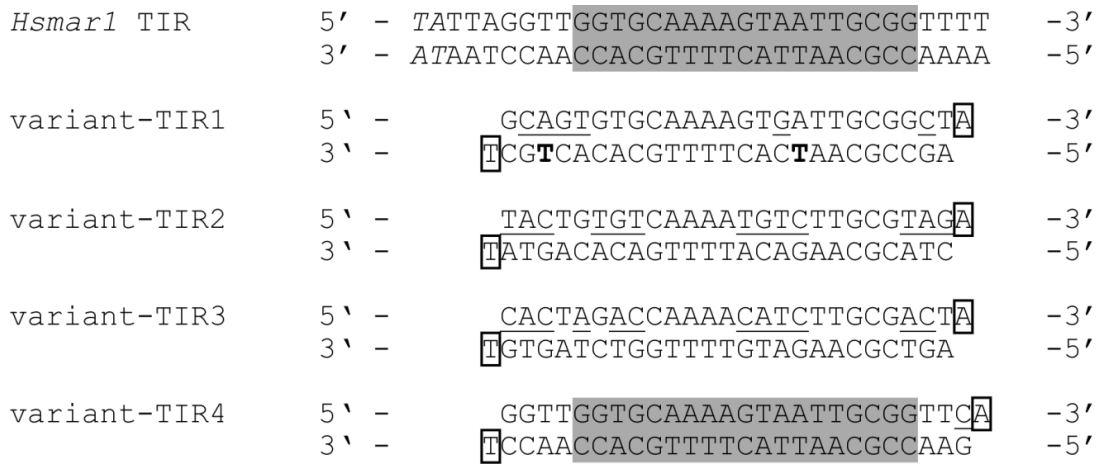

Figure 13. *Hsmar1* TIR based DNA sequences used for crystallization.
The left end *Hsmar1* TIR sequence is shown on the top. The italic TA is the characteristic target site duplication in Tc1/*mariner* transposon superfamily. Four variant-TIR sequences were used for crystallization trials. The core 19-bp binding site is in gray. The underlined bases are mutated from the *Hsmar1* TIR sequence. In variant-TIR1, the bold Ts represent positions for which T was replaced by 5-BrdU in the Br-DNA duplexes. The overhang nucleotides in boxes are designed for facilitating crystal packing.

Protein construct	329-440 (wt)	329-440 (C381S)	329-440 (C381S)	316-440 (C381S)	329-440 (C381R)
DNA	TIR 2	TIR 2	variant-TIR 1	variant-TIR 1	TIR 2
Crystallization condition	0.2 M lithium sulfate 0.1 M Bis-Tris pH5.5 25% w/v PEG 3350	25% poly (acrylic acid sodium) 5100 0.10M HEPES pH7.5 0.02M MgCl2	0.01 M magnesium acetate 0.05 M MES pH5.6 2.5 M ammonium sulfate	0.10 M magnesium formate 15% PEG3350	0.025 M magnesium sulfate hydrate 0.05 M Tris-HCl pH 8.5 1.8 M ammonium sulfate
Space group	N/A	C2	N/A	$C222_1$	$C222_1$
Cell parameters	N/A	a=78.7 Å b=168.6 Å c=74.9 Å β=108.35°	N/A	a=72.23 Å b=164.39 Å c=67.96 Å	a=74.727Å b=168.760 Å c=72.191 Å

Table 3. Summary of the initial crystals. N/A: not available

In a parallel effort, we screened the 329-440(C381R) DBD complexed with both TIR and variant-TIR sequences. Crystals were obtained for a complex with TIR2 from Hampton Research Natrix D9, which contains 0.025 M magnesium sulfate hydrate, 0.05 M Tris hydrochloride pH 8.5, and 1.8 M ammonium sulfate (Figure 12 F). Crystals diffracted to 4 Å and belong to space group $C222_1$ with cell dimensions of a=74.727Å, b=168.760 Å, and c=72.191 Å. No promising crystals were obtained from 329-440(C381R) with variant-TIR 1 complex.

B. Considerations in phasing strategies

Although the sequence of the SETMAR's DBD is 30 % identical to that of Mos1, the DNA sequence that it recognizes makes up 54% of the mass of the complex and shares no sequence similarity to the Mos1 sequence. Thus, we pursued two single anomalous dispersion (SAD) phasing strategies for selenomethionine-labeled protein (SeMet SAD)

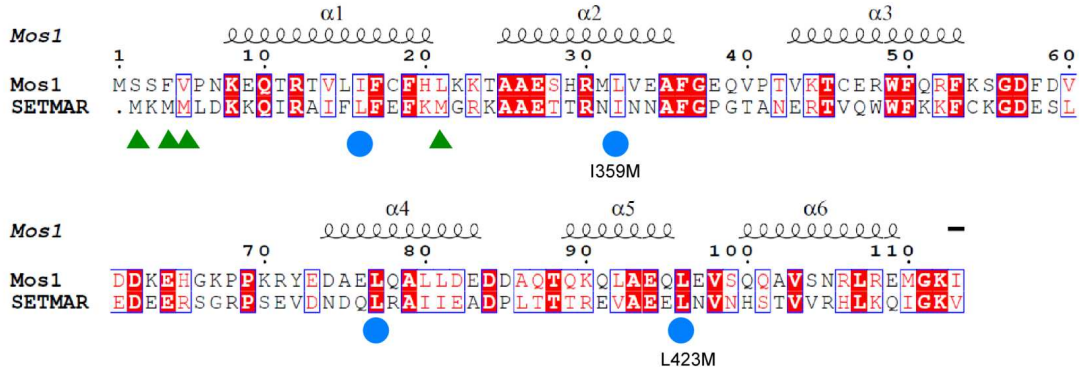

Figure 14. The pairwise sequence alignment of Mos1 transposase and SETMAR DNA-binding domains, 1-113 and 329-440, respectively.
The intrinsic Met residues are marked by green triangles. Residues substituted with Met for phasing by SeMet SAD are highlighted by blue dots. Secondary structure elements of Mos1 transposase DNA-binding domain (PDB ID: 3HOT) are shown above the alignment. (The figure was created with ESPript 3.0)

and brominated DNA (Br SAD). For experimental SeMet SAD phasing purposes, several Met substitutions were introduced to ensure a robust Se anomalous signal. Since three out of the four intrinsic Met residues are clustered on the N-terminus and are likely to be disordered, we sought to improve the Se signal by increasing the number of ordered SeMet residues within the protein. Through a sequence comparison with Mos1, we identified amino acid residues that might be well-ordered and tolerate substitution to

Met within both predicted HTH motifs in SETMAR (Figure 14). Desirable sites for introduction of Met include those that are conserved as hydrophobic residues Leu, Met, or Ile in both sequences. Four doubly mutated constructs were made including the following pairs: L343M and L404M, L343M and L423M, I359M and L404M, and I359M and L423M. The double mutant protein constructs were introduced into the 329-440(C381R) plasmid. Of these, SETMAR (C381R)(I359M)(L423M) was stably expressed in sufficient quantities for crystallization experiments. With the introduction of two additional Mets, there are a total of six possible Se sites in the SeMet-labeled protein (Figure 14).

For Br SAD experimental phasing, two thymidine sites (dT) were replaced with Br-dU in two strands for TIR2 and within a single strand for variant TIR1, with the criteria that those sites were not symmetrically positioned within the oligonucleotide and would not be expected to be in direct contact with the protein based on the prediction (Figure 15). Unexpectedly, crystals grown for DBD complexes with brominated oligonucleotides were uniformly of better quality than those grown for the corresponding complex with the natural DNA sequences. As a consequence, all of the crystals used for phasing experiments contained brominated oligonucleotides. Although Br SAD data were collected for DBD 316-440(C381S) complexed with variant brominated-TIR1 and DBD 329-440(C381R) complexed with brominated TIR2, attempts to identify Br sites were not successful and thus no phasing information was obtained.

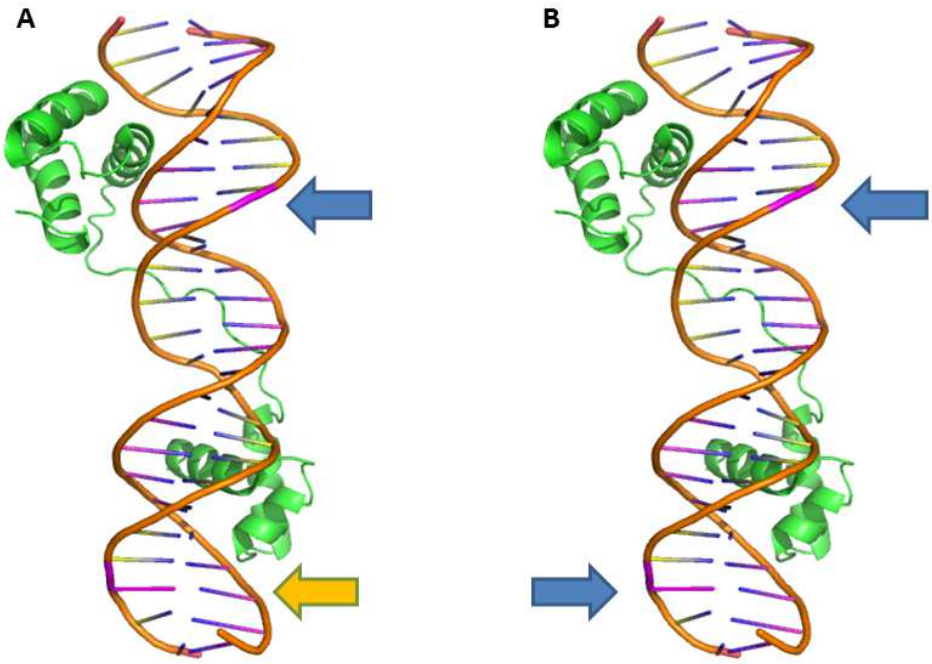

Figure 15. Predicted BrdU replacement sites based on Mos1 DNA-binding domain complex with Mos1 TIR model (PDB ID: 3HOT).
(A) Predicted model for TIR2 DNA. Two sites are located in different strands, shown as blue and yellow arrows. (B) Predicted model for variant-TIR1 DNA. Both sites are located in the same strand, shown as blue arrows. For DNA sequences, see Figure 11 and Figure 13.

C. Low resolution Se SAD phasing

Two datasets were collected at the Advanced Photon Source beamline 23-ID-D for SeMet derivatized DBD 329-440(C381R)(I359M)(L423M) complexed to brominated TIR DNA, one to 3.75 Å and the other to 3.24 Å resolution (Table 4). Attempts to identify

Data set	C10	C11	Merged
X-ray source	GM/CA 23-IDB	GM/CA 23-IDB	GM/CA 23-IDB
X-ray detector	MARCCD	MARCCD	MARCCD
Wavelength(Å)	0.97945	0.97945	0.97945
Space group	$C222_1$	$C222_1$	$C222_1$
a, b, c (Å)	72.41, 167.01, 66.61	71.63, 166.29, 65.40	72.02, 166.65, 66.01
Resolution range (Å)	47.04--3.75	83.14—3.24	46.71—4.17
Completeness (%)	98.0	99.6	99.9
Redundancy	11.3	11.0	21.3
Mean I/σ(I)	13.0	10.8	20.5
R_{pim} or R_{meas}	0.050	0.067	0.038

Table 4. Data statistics of 329-440(C381R)(I359M)(L423M) complex with TIR DNA.

Se sites from either dataset alone were not successful. Therefore, using the CCP4 (Winn et al. 2011) program BLEND (Foadi et al. 2013), the two SeMet SAD datasets were merged producing a 4.17 Å dataset (Table 4). Three of the possible six sites were identified using the program AutoSol as implemented in PHENIX (Adams et al. 2010), namely I359M, L423M and the intrinsic M348. These sites were used to phase an electron density map at 4.17 Å. The electron density of the DNA phosphate backbone was identified in this map, and an initial model for the DNA phosphate backbone and a polyalanine model for the two HTH motifs were built (Figure 16).

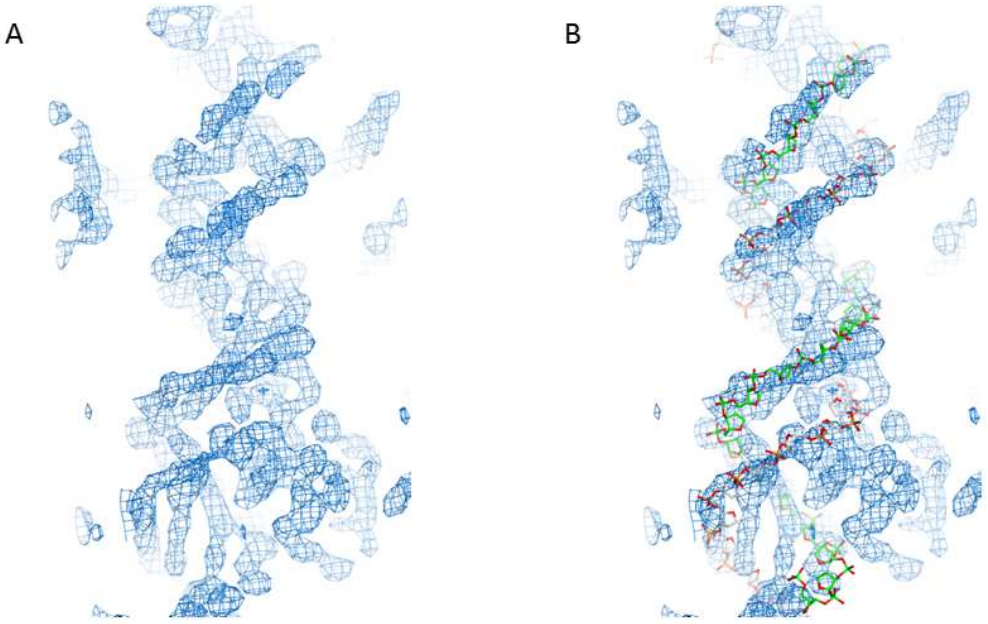

Figure 16. The 4.17 Å experimental electron density map contoured at 2.0 sigma, using merged SeMet SAD data set.
(A) SAD map shows DNA backbone density. (B) An initial model of the DNA phosphate backbone fits into the SAD map contoured at 2 sigma.

D. Crystal structures of SETMAR bound to two different DNA sequences reveal a conserved set of interactions

Having completed the above phasing experiments, we optimized the crystallization of SeMet-derivatized 329-440(C381R) (I359M) (L423M) complexed to TIR DNA (the TIR complex) and ultimately obtained crystals that diffracted to higher resolution. An initial structure of the TIR complex was determined by Se-SAD phasing at 2.66 Å, figure of merit 0.38 for initial phasing, (Figure 17 A) with well-defined electron density for residues 330-437 and the entire DNA duplex. Anomalous difference Fourier analysis was used to confirm the location of five Se-Met sites in the model (Figure 17 B). This

structure was then used as the search model to determine a higher resolution TIR complex structure and the variant TIR complex structure by molecular replacement. In this thesis, we present the crystal structures of SETMAR DBD bound to two different 25-mer DNA duplexes: DBD 329-440(C381R) complexed to TIR DNA (the TIR complex) at 2.37 Å resolution and DBD 316-440(C381S) complexed to variant-TIR DNA (the variant TIR complex) at 3.07 Å.

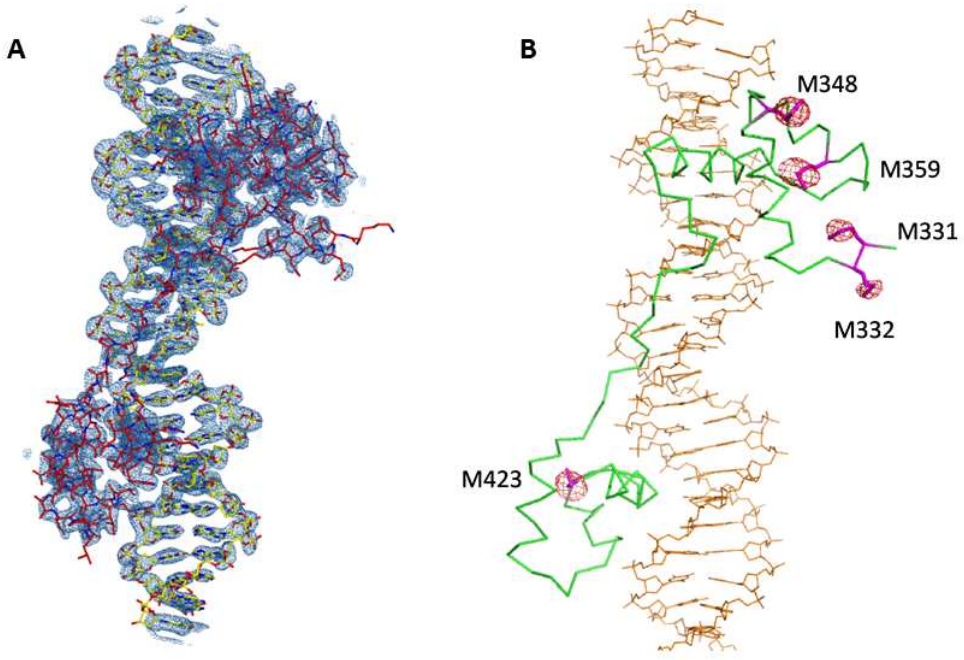

Figure 17. Se-SAD phasing.
(A) The experimental electron density map of TIR complex, from selenium single wavelength anomalous diffraction (Se-SAD) phasing using data to 2.66 Å (blue mesh, contoured at 1.6 sigma). The electron density map is superimposed on the refined model. The protein backbone atoms are shown in red and the DNA backbone atoms are shown in yellow. (B) Anomalous difference Fourier map of the SeMet-labeled TIR complex, superimposed with the backbone trace of the refined model (green for protein and orange for DNA). The map is contoured at 4 sigma (red). (SeMet 331, 332, 348, 359, and 423 are displayed as ball-and-stick side chain models (magenta color).

	TIR complex (Se-SAD)	TIR complex (High Res.)	Variant-TIR complex
PDB ID	-	5KWT	5KWU
Data Collection			
Space group	$C222_1$	$C222_1$	$C222_1$
Cell dimensions			
a, b, c (Å)	70.09, 166.04, 66.09	70.98, 166.17, 66.05	72.10, 164.21, 68.16
α,β,γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Wavelength (Å)	0.97938	0.97938	0.91922
Resolution (Å)	29.42—2.66	27.70—2.37 (2.46—2.37)	82.11---3.07 (3.11—3.07)
R_{pim}	0.027	0.024 (0.256)	0.050
Mean I/σ(I)	18.4	21.9 (2.3)	11.5 (1.87)
Completeness (%)	99.5	99.7 (99.1)	97.7 (91.3)
Redundancy	7.7	4.3 (4.0)	8.3 (7.5)
Refinement			
Resolution (Å)	-	27.70---2.37	82.11---3.07
No. reflections	-	16176	
R_{work} / R_{free} (%)	-	21.10 / 23.50	22.45 / 25.44
No. atoms	-		
Protein	-	884	870
DNA	-	1060	1060
Water	-	13	0
B factors	-		
Protein	-	73.23	126.88
DNA	-	89.69	158.19
Water	-	71.07	-
r.m.s. deviation	-		
Bond lengths (Å)	-	0.013	0.011
Bond angles (°)	-	1.34	1.60

Table 5. Data collection and refinement statistics.

The structures reveal a dimeric complex created by crystallographic symmetry in which each DBD comprises two HTH motifs connected by a 17 aa linker (residues 384-400) containing two AT hook elements bound to a 25mer DNA duplex (Figure 18). Each HTH motif contains three α-helices with the third α-helix, the recognition α-helix, in each HTH motif (α3 or α6) bound to the major groove of the DNA (Figure 18). Interactions with the minor groove of the DNA involve the AT hook elements within the linker between α3 and α4 helices. The two HTH domains differ in size with HTH1 comprising 47 and HTH2 37 residues and therefore in structure with a rmsd of 1.7 Å for superpositioning of 34 Cα atoms. The larger of the two HTH motifs, HTH1, dimerizes through interactions of residues F344, F363, and I341 from α1 and α2 helices burying 1610 $Å^2$ in the interface (Figure 19). In the smaller HTH2 domain, α4 and α5 are both shorter than corresponding α1 and α2, and the loop connecting α5 and α6 is 7 residues shorter than the comparable loop in HTH1.

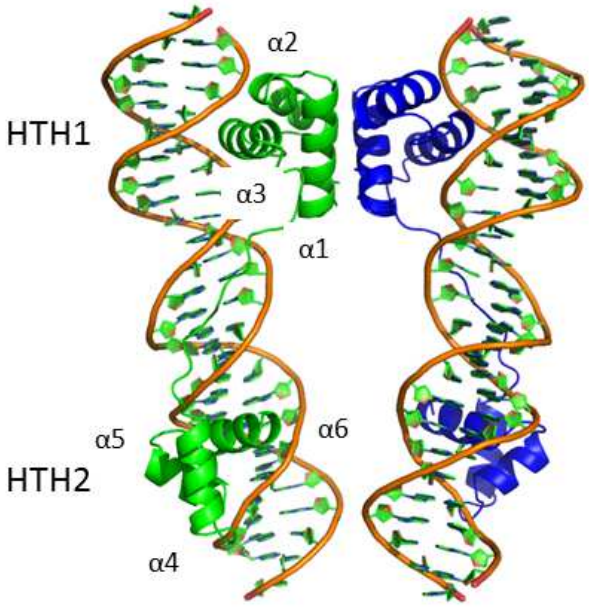

Figure 18. The SETMAR DNA-binding domain is a dimer.
Each monomer has HTH1 and HTH2 motifs that bind the major-groove of DNA. A linker region containing two AT-hook motifs between them binds the minor-groove of DNA. HTH1 motifs from each of the monomers form a dimerization interface.

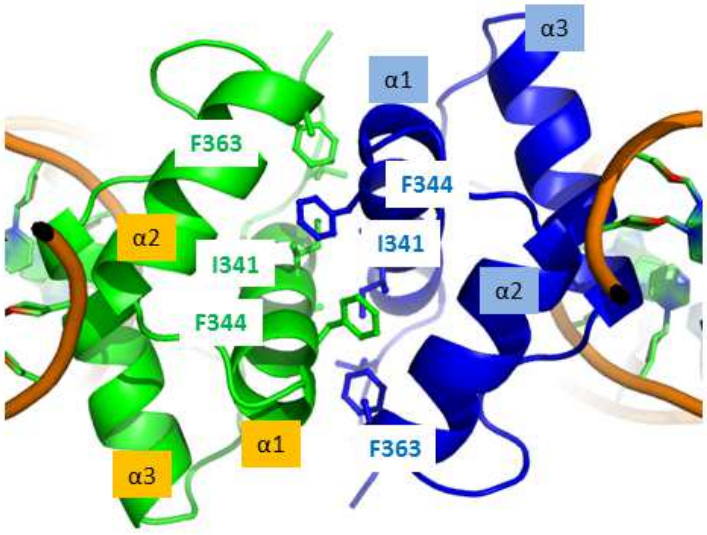

Figure 19. Protein-protein interface of HTH1 motifs.
Hydrophobic residues involved in the interface are shown as sticks. Two representative residues of the hydrophobic cluster, I341, F344, and F363, are labeled.

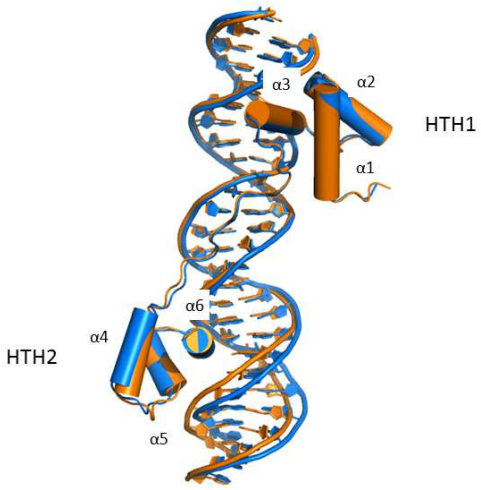

Figure 20. Superimposition of TIR and variant-TIR complex.
HTH1 motifs superimpose quite well in the two complexes. In the HTH2 motif, the alpha helix 4 is shifted slightly away from the DNA in the variant-TIR complex (marine blue), compared to that of the TIR complex (orange). Alpha helices are shown as cylinders.

Overall, the structures of SETMAR DBDs bound to the TIR and variant DNA sequences are similar with an rmsd of 0.75 Å for superpositioning of 107 Cα atoms but differ specifically in the contacts made between HTH2 and the DNA (Figure 20). There are a total of 32 hydrogen bonding interactions between the protein and DNA in the TIR complex (Figure 21), 18 involving the phosphodiester backbone, and 20 in the variant-TIR complex structure, 13 of which are to the phosphodiester backbone (Figure 22). The structures of the DNA also differ as expected due to differences in sequence (Figure 20) with the TIR DNA being primarily B-form, while that of the variant DNA includes several regions that deviate from this form, nucleobase pairs B17-B20/C9-G6 and B9-B12/C17-C14 (with B and C referring to chain IDs). The second region differs in

sequence from the TIR at position B10 with a T:A to C:G nucleobase-pair substitution and interacts with the first AT hook residue R392.

In both structures, R371 in HTH1 motif makes a nucleobase-specific interaction within the major grove with atoms NE and NH2 forming two hydrogen bonds to O6 and N7 of G5, B chain (Figure 23). In HTH2, four residues (R417, H427, S428, and R432) make nucleobase-specific contacts with DNA in the major groove in the TIR complex while only two of these, R432 and S428, make nucleobase-specific contacts in the variant TIR complex (Figure 24). In the TIR complex, R417 forms two hydrogen bonds between its NH1 and NH2 atoms with N7 atom and O6 atom of C chain G5. H427 NE2 hydrogen bonds with O6 atom of C chain G6, S428 OG with atom N4 of B chain C18, R432 NH1 and NH2 with O6 atoms of B chain G17 and C chain G8, respectively, and R432 NH2 with N4 atom of C chain C9. However, in the variant TIR complex, only S428 and R432 hydrogen bond to DNA (Figure 24).

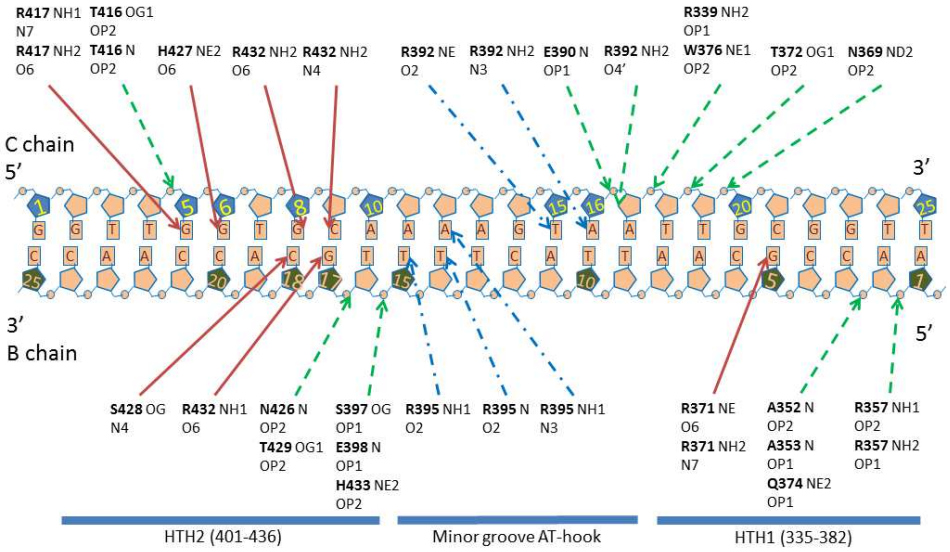

Figure 21. Schematic diagram of hydrogen-bonding interactions between protein and DNA in the TIR complex.

Red arrows are major groove base-specific contacts. Blue dotted arrows are minor groove contacts. Green broken arrows are interactions between protein and DNA backbone.

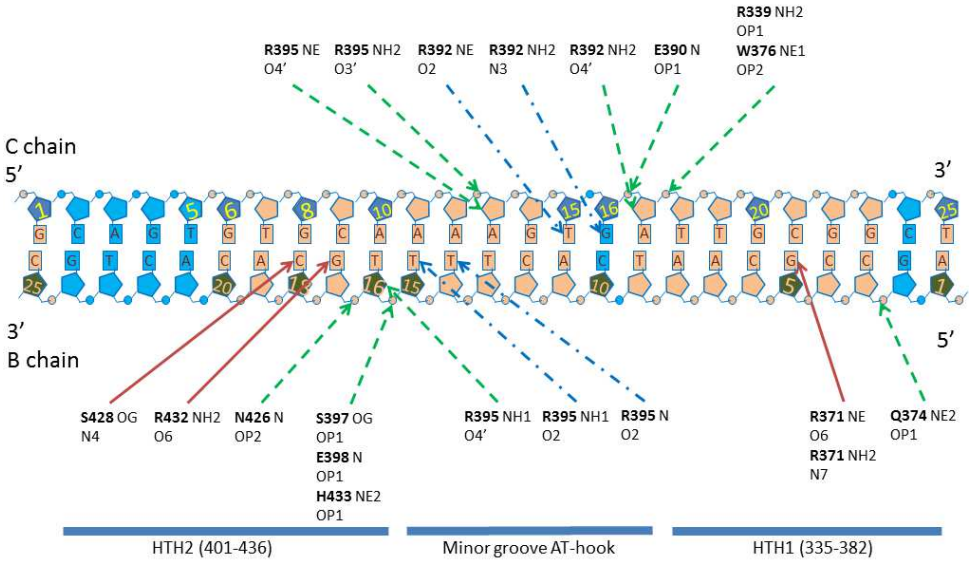

Figure 22. Schematic diagram of hydrogen-bonding interactions between protein and DNA in the variant-TIR complex.

Nucleotides different from the TIR DNA sequence are shown in blue. Red arrows are major groove base-specific contacts. Blue dotted arrows are minor groove contacts. Green broken arrows are interactions between protein and DNA backbone.

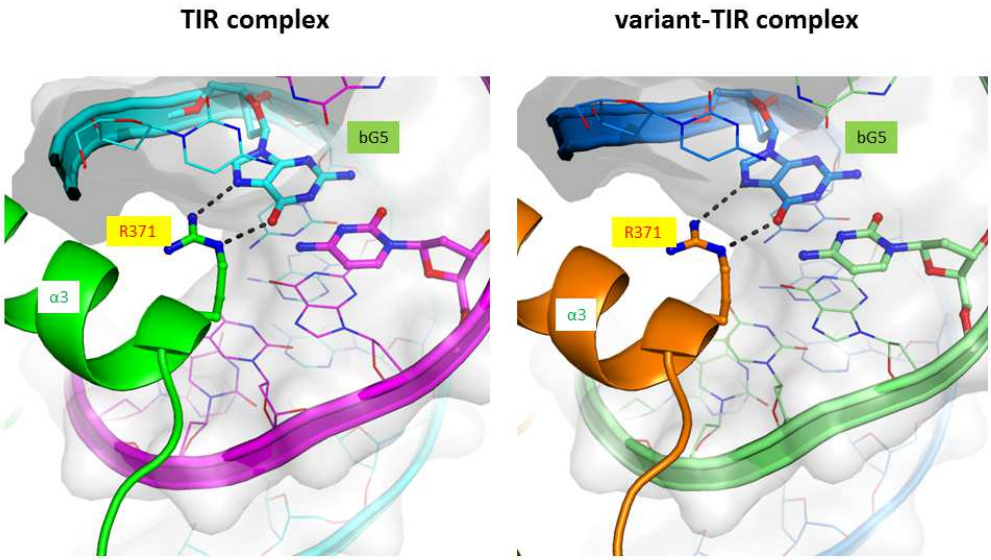

Figure 23. Close-up view of base-specific contacts made by Arg-371 in HTH1 motif. R371 makes two hydrogen bonds (broken lines) with the guanine-5 from chain B (labeled as bG5). Both TIR (left) and variant-TIR complex (right) show a similar interaction. Protein is renderer as a ribbon diagram. DNA is shown as line and surface rendering. Atoms oxygen and nitrogen are shown red and blue, respectively. Key residue side chain and bases are shown as ball-and-stick models.

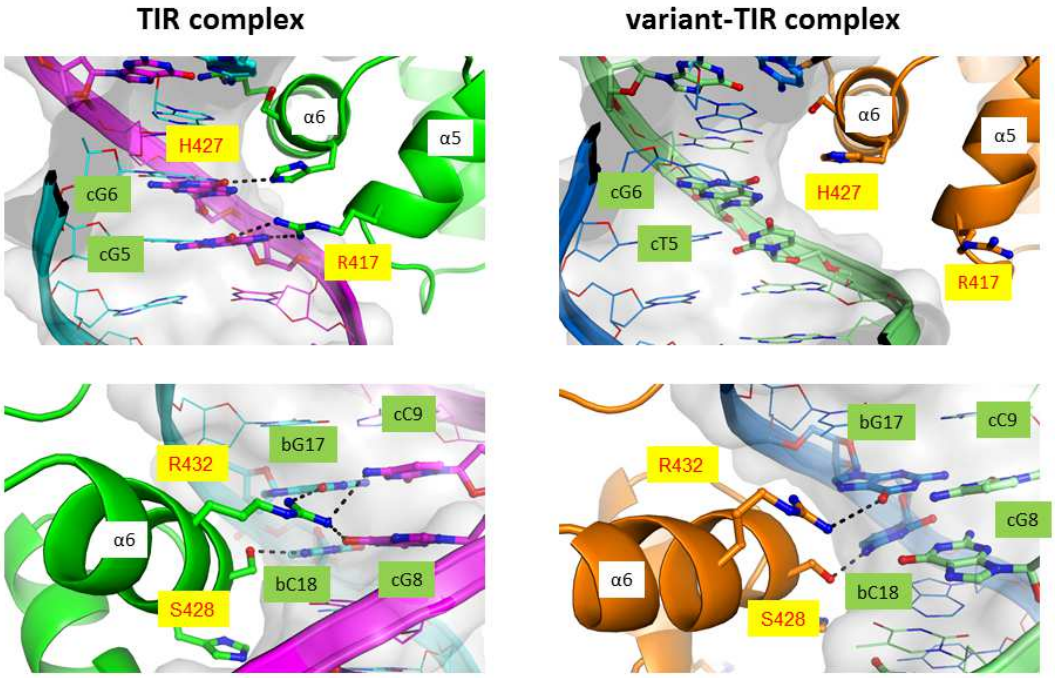

Figure 24. Close-up view of key differences in base-specific contacts in HTH2 motifs of TIR and variant-TIR complex.
In the TIR complex (left column), key residues (R417, H427, S428, and R432) make 7 hydrogen bonds with DNA nucleobases. While in the variant-TIR complex (right column), two base-specific contacts are made by S428 and R432. Note that in the variant-TIR complex, the nucleobase in position 5 of the DNA chain C is T5 (labeled as cT5), instead of guanine in the TIR complex (labeled as cG5). Hydrogen bonds are shown in black broken lines. The TIR complex is on the left, variant-TIR complex on the right.

SETMAR's DBD includes two AT-hooks, one with a non-canonical sequence ERS (391-393) and the second a canonical GRP (394-396) (Aravind and Landsman 1998), in which the central Arg is bound to the minor groove of the DNA. Within this contact region in both structures, R392 is stabilized through an interaction with D389, which also interacts with R339. NH2 of R392 hydrogen bonds to N3 of either C chain A16 in the TIR or G16 in the variant TIR and NHE to O2 of T15 C chain (Figure 25). Note that in this case, the sequence variation between TIR and variant TIR does not affect the minor groove

interactions. However, in the TIR complex, R395 is oriented with its guanidinium group such that NH1 points into the groove and hydrogen bonds with N3 of A12 (C chain) and with O2 of T15 (B chain) in a classic bifurcated type of interaction typically seen in narrow AT-rich minor grooves. (Figure 26) R395 is positioned by a hydrogen-bonding interaction of its NHE with OE2 of E398. Although the minor groove widths in both structures are similar in this region, the guanidinium group of R395 in the variant TIR complex is slightly rotated relative to the conformation in the TIR complex and forms a single hydrogen bond between NH1 and O2 of T15 (B chain) (Figure 26). In both structures, the AT-hooks also make several backbone interactions (Figure 21 and Figure 22).

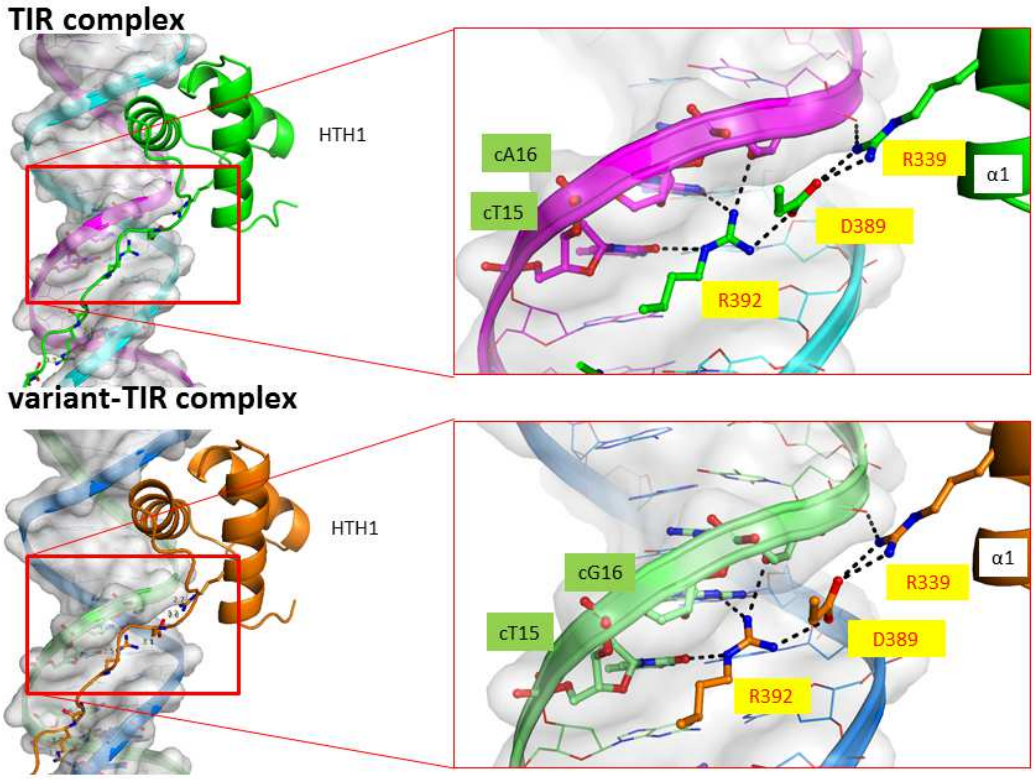

Figure 25. The interface between the N-terminal linker region and the DNA minor groove.

Arg-339 and Asp-389 orientate Arg-392 to make contacts with DNA bases in the minor groove. Both TIR (top) and variant-TIR (bottom) complexes have similar contacts. Note that in the variant-TIR complex, the nucleobase in the position 16 of the DNA chain C is guanine (labeled as cG16), instead of adenine in the TIR complex (labeled as cA16). Hydrogen bonds are shown in black broken lines. Protein is rendered as a ribbon diagram. DNA is shown as line and surface rendering. Atoms oxygen and nitrogen are shown red and blue, respectively. Key residue side chains and bases are shown as ball-and-stick models.

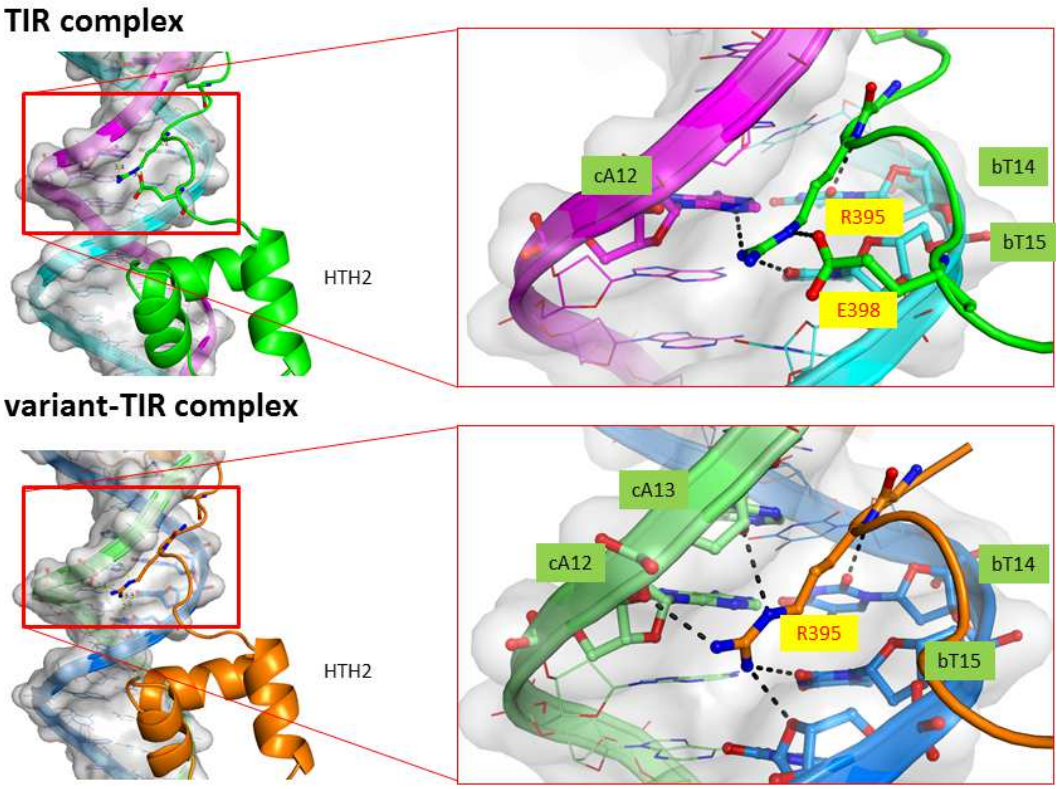

Figure 26. Differences in DNA contacts in the interface between the C-terminal linker region and the DNA minor groove.

The guanidinium groups of Arg-395 adopt different conformations in two complexes, resulting in different contacts between protein and DNA. In a classic bifurcated type of interaction mode, NH1 of Arg-395 hydrogen bonds with two bases: adenine 12 at chain C (labeled as cA12) and thymine 15 at chain B (labeled as bT15) (close-up view on top). In the variant-TIR complex (bottom, close-up view), the guanidinium group of R395 is slightly rotated relative to that in the TIR complex and forms a single hydrogen bond between NH1 and O2 of thymine 15 at chain B (labeled as bT15). Note that in the TIR complex (top), OE2 of Glu-398 makes hydrogen bond with NE atom of Arg-395. Hydrogen bonds are shown in black broken lines. Protein is rendered as a ribbon diagram. DNA is shown as a line and surface rendering. Atoms oxygen and nitrogen are shown red and blue, respectively. Key residue side chains and bases are shown as ball-and-stick models.

E. SETMAR binds TIR and variant TIR DNA with similar affinity

In order to characterize the DNA-binding activity in the context of full-length SETMAR, we conducted fluorescence anisotropy (FA) assays using rhodamine-labeled DNA probes. A K_D value of 53 $\pm$ 4 nM was measured for binding of SETMAR to an *Hsmar1* TIR DNA probe as compared to 91 $\pm$ 7 nM for the variant DNA sequence used for crystallization (Figure 27). The reduction in binding affinity to the variant sequence is consistent with fewer nucleobase-specific hydrogen bonds formed between HTH2 and the DNA in this complex as compared to the TIR sequence. However, it is notable that the reduction in affinity is less than two-fold despite significant differences in the DNA sequence and its recognition pattern by HTH2.

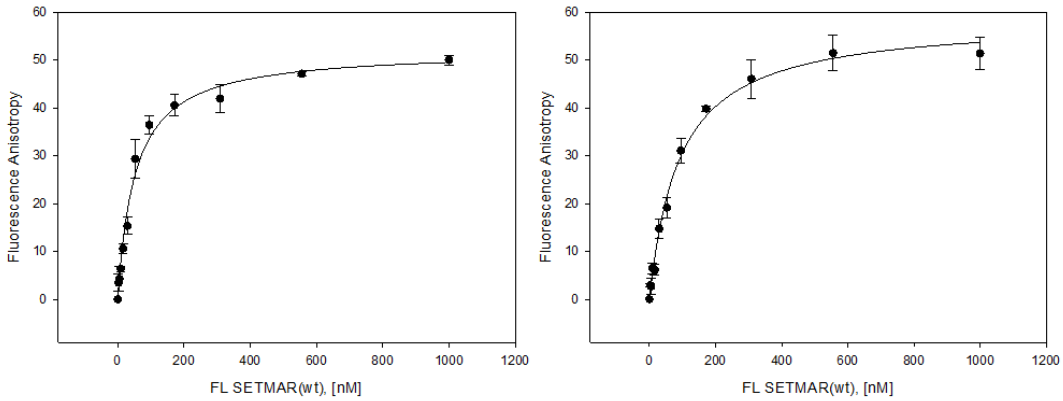

Figure 27. Fluorescence anisotropy (FA) assays characterize DNA-binding affinity of full-length SETMAR (FL SETMAR) and DNA.
Rhodamine-labeled TIR or variant-TIR probes (10 nM) were titrated with increasing amount of SETMAR protein. Binding curves were fitted by fluorescence anisotropy (FA) against protein concentration. Kd value for TIR probe (Left) binding is 53 ± 4 nM. For variant-TIR probe (Right), Kd value is 91 ± 7 nM. Experiments were conducted with triplicate reading for three independent assays.

Name	Sequence (5'—3')
Hsmar1 TIR	TTAGGTTGGTGCAAAAGTAATTGCGGTT
Mos1 TIR	TCAGGTGTACAAGTATGAAATGTCGTTT
Hsmar1 TIR without sequence-specific binding sites	TTAGGTT**AA**T**AT**AAAAGTAATTG**T**GGTT
Hsmar1 variant TIR	TTAG**CAGT**GTGCAAAAGT**G**ATTGCGG**C**T
No_G variant TIR	TTAG**CAGT**GTGCAAAAGT**G**ATTG**T**GG**C**T
No_CG variant TIR	TTAG**CAG**GGT**AT**AAAAGT**G**ATTGCGG**C**T
No_CGG variant TIR	TTAG**CAGT**GT**AT**AAAAGT**G**ATTG**T**GG**C**T

Table 6. DNA oligos for competition assay.
The underlined bases are mutated from the *Hsmar1* TIR sequence.

To assess the importance of specific nucleobases that interact with the HTH binding motifs, we examined the ability of mutant TIR or variant sequences to compete with labeled TIR or variant TIR DNA for binding to SETMAR by fluorescence anisotropy (See Table 6). First, we performed a competition assay using unlabeled DNA to compete off the fluorescently-labeled DNA probe in the protein-DNA complex. By replacing all nucleobase-specific interacting guanines (G) to adenines (A), the resulting G to A mutant TIR DNA was unable to compete off the probe, acting like the negative control Mos1 TIR DNA (Figure 28 A). Similarly, an oligonucleotide in which guanines that interact with SETMAR in the variant TIR complex were progressively replaced with adenines failed to compete effectively with the fluorescently-labeled variant DNA sequence (Figure 28 B).

These results confirm that the nucleobase-specific interactions are important for DNA-binding activity of SETMAR.

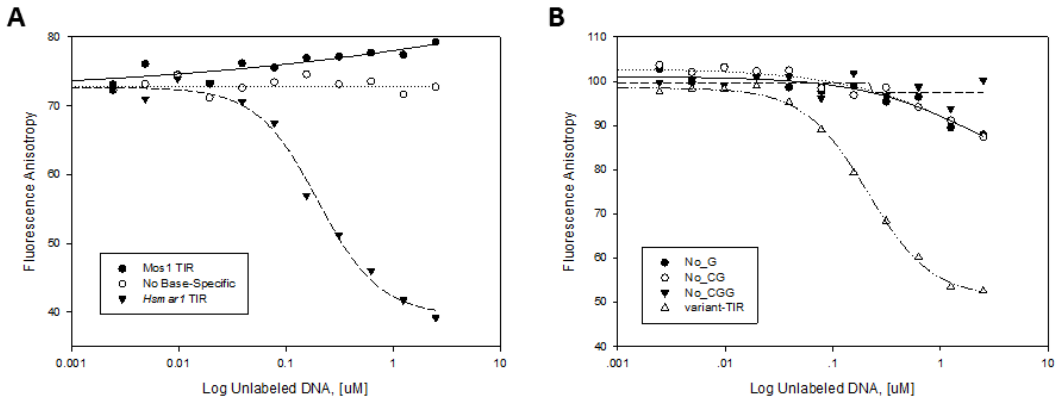

Figure 28. Competition assays of SETMAR with DNA probes.
(A) Competition assay using various non-fluorescence labeled DNA sequences. Non-fluorescence labeled *Hsmar1* TIR DNA (filled triangle) was titrated to compete off the bound TIR DNA probe from a complex, showing a curve with decreasing fluorescence polarization. Mos1 TIR (filled circle), a non-cognate DNA sequence for *Hsmar1* transposase, was unable to compete off the bound TIR DNA probe, serving as a negative control. As a consequence of substitutions of all key nucleotides involved in major groove interactions from guanine to adenine, a mutant *Hsmar1* TIR DNA (empty circle) had no competition capability at all, behaving like the negative control. (B) Competition assay using non-fluorescence labeled mutant forms of variant-TIR DNA. Non-labeled variant-TIR DNA (empty triangle) was titrated to compete off the bound variant-TIR DNA probe from a complex, showing a curve with decreasing fluorescence polarization. Variant-TIR sequences bearing mutations in the key protein-recognition nucleotides were unable to compete off the probe. R371-contacting guanine was mutated to adenine (filled circle). S428 and R432 contacting CG dinucleotide was mutated to TA dinucleotide (empty circle). A mutant sequence carrying all above mentioned substitutions (filled triangle) had no competition activity at all.

To determine the contributions of specific amino acids within HTH1 and HTH2 to the DNA-binding activity of SETMAR, we substituted key residues with Ala and measured their binding affinity for the *Hsmar1* TIR DNA probe. Residues selected for analysis include R371, S428, and R432 which make similar hydrogen bonding interactions in both TIR and variant sequences. K_D values for binding of wild-type, R371A, S428A, and R432A to TIR DNA are 42, 521, 485, and 302 nM, respectively (Figure 29), consistent with an important role for these residues in this protein-DNA interaction.

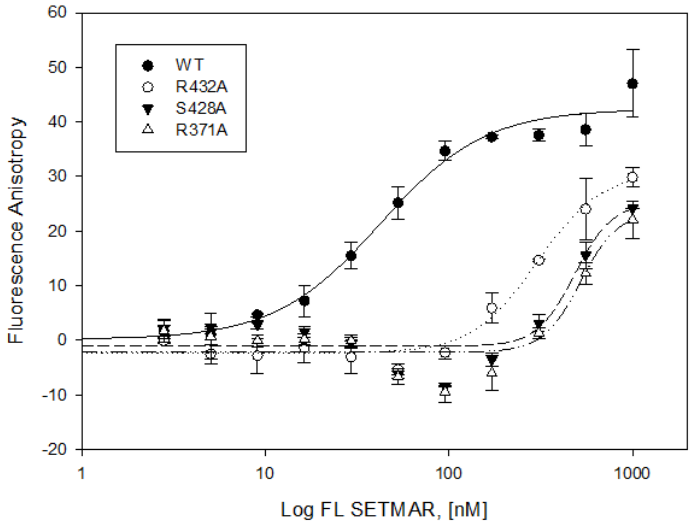

Figure 29. Mutations in key amino acid residues decrease DNA-binding affinity of SETMAR. DNA-binding assays were conducted as Figure 27. The binding affinity of the mutants decrease approximately 10-fold compared to that of wild-type SETMAR. Experiments were conducted with triplicate reading in three independent assays.

F. SETMAR binds DNA in cell-based assays

SETMAR's DBD comprising two HTH motifs is structurally very similar to the paired homeodomain family of transcription factors including PAX6, for which a DNA-bound crystal structure has been reported (Xu et al. 1999). Members of Tc1/*mariner* transposon superfamily share a common ancestor with PAX proteins (Breitling and Gerber 2000). Given that the DBD of SETMAR is derived from a Tc1/*mariner* superfamily transposon, we examined the possibility that SETMAR evolved to regulate gene expression. To test this hypothesis, we conducted luciferase reporter assays. Five tandem repeats of the *Hsmar1* TIR sequence were inserted upstream of an SV40 promoter of the pGL3-Promoter Vector (Promega, Madison, WI), which we refer to as the pGL3-promoter-5xTIR. The luciferase reporter vector, pGL3-promoter-5xTIR, was cotransfected with the overexpression vector, pFLAG-CMV4-SETMAR(wt), into HEK293T cells, which have no detectible SETMAR expression (Williamson et al. 2008, De Haro et al. 2010). Compared to a control experiment lacking the overexpression vector, SETMAR(wt) decreased the relative luciferase activity (Figure 30) by approximately 30%. However, SETMAR(R371A), a DNA-activity deficient mutant (Figure 29), was unable to repress the luciferase gene expression. Compared to wild-type protein, the R371A mutation doesn't affect the expression of the SETMAR, based on western-blot result (Figure 30 inset).

To confirm that the observed repression in the luciferase reporter assay results from binding of SETMAR to the TIR sequence within the plasmid, we conducted a ChIP assay in cells: in this case we measured SETMAR binding to TIR sequences embedded in

plasmid reporter DNA. Using the same vectors and cells as the luciferase reporter assay, FLAG-tagged SETMAR bound to DNA was immunoprecipitated using Anti-FLAG agarose beads, and TIR DNA was quantitated by qPCR resulting in a 3-fold higher signal than that of an IgG negative control group (Figure 31). Together with the luciferase results for the R371A mutant, this result is consistent with repression of luciferase activity by SETMAR resulting from TIR-specific binding activity.

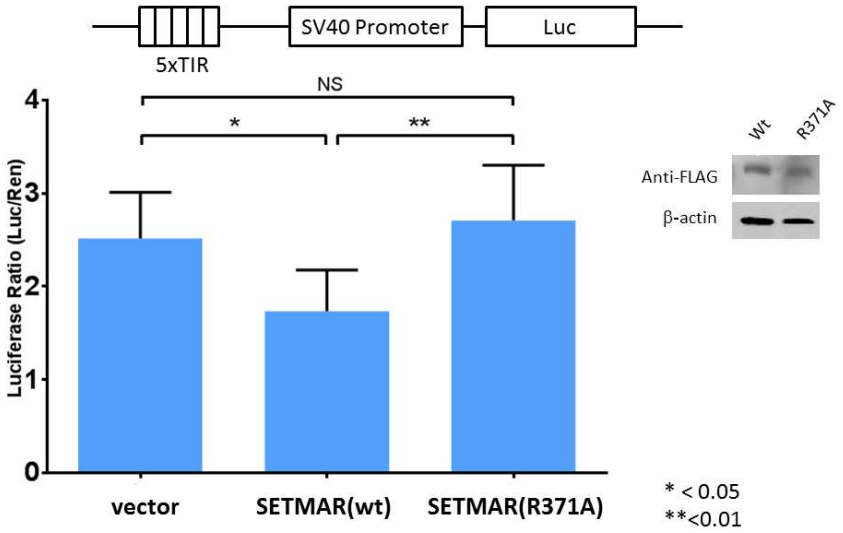

Figure 30. SETMAR represses transcription in a luciferase reporter assay.
The pGL3-promoter-5xTIR (5xTIR) were co-transfected with pFLAG-CMV4-SETMAR(wt) or pFLAG-CMV4-SETMAR(R371A) into HEK293T cells. Error bars represent S.E.M. Samples significantly different between groups are indicated by asterisks (*: $p<0.05$, **: $p<0.01$). NS: non-significantly difference. Student t-test was used. Inset is the western-blot of FLAG-tag SETMAR(wt) and FLAG-tag SETMAR(R371A), showing that the residue substitution does not affect the protein expression level and the stability.

SETMAR has been demonstrated to promote repair and restart stalled replication forks. From the cisRED database (Robertson et al. 2006), we identified a potential SETMAR binding site within the promoter region (-113 to -93 bp region) of *TOPBP1*: *5'*-**GGTT<u>GG</u>C<u>GC</u>G<u>AAAGT</u>**CGG**T**T**<u>C</u>** -3'(bold letters are identical nucleobases in the *Hsmar1* TIR sequence, underlined letters indicate nucleobases involved in specific contacts with SETMAR or in the case of the final C, the complementary nucleobase). This sequence retains all of the nucleobases that are recognized in the major groove by SETMAR. TOPBP1, DNA topoisomerase II-binding protein 1, plays a role in the rescue of stalled replication forks and checkpoint control (Emmons et al. 1980, Emmons et al. 1983, Jurka et al. 2005). To determine whether SETMAR binds to the *TOPBP1* promoter region, we performed a ChIP experiment with overexpression FLAG-tag SETMAR(wt) in HEK293T cells and found a 1.5-fold enrichment of bound SETMAR to the *TOPBP1* promoter region

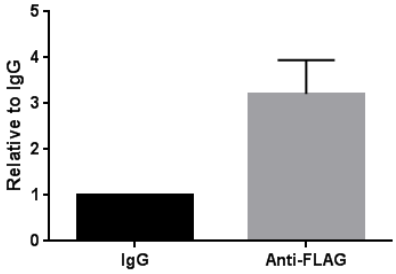

Figure 31. SETMAR binds TIR sequence in cells.
Using the same vectors and cells as Figure 30, FLAG fusion SETMAR is cross-linked to 5xTIR vector DNA and then immunoprecipitated by anti-FLAG M2 affinity agarose. Purifed DNA was analyzed by quantitative Real-Time PCR, using primers flanking 5xTIR sequence in the vector. The amount of immunoprecipitated DNA is represented as signal relative to IgG negative control. Data were presented as mean +/- SEM for three independent experiments ($p<0.05$).

as compared to the IgG negative control (Figure 32). This result is consistent with binding of SETMAR to the identified site within the *TOPBP1* promoter.

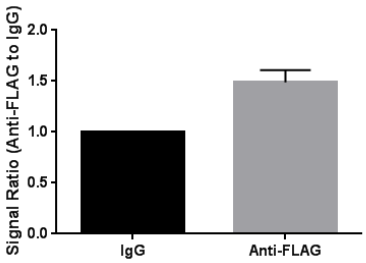

Figure 32. Binding of FLAG-tag SETMAR to the promoter region of *TopBP1* gene in HEK293T cells is detected by ChIP assay.

FLAG-tag SETMAR cross-linked on the chromatin in HEK293T cells was immunoprecipitated by Anti-FLAG M2 affinity agarose gel. The SETMAR-binding site in the promotor region of *TopBP1* gene in the precipitated chromatin was detected by qPCR using specific primers for amplification of the -188 to -46 bp region. Mouse IgG-conjugated agarose was used as a negative control. Signals from anti-FLAG group were normalized to the IgG group. Data were presented as mean +/- SEM for three independent experiments ($p<0.05$).

DISCUSSION

A. Structural basis of the sequence-specific binding activity of *Hsmar1*

During the life cycle of a DNA transposon, one of the key steps for the encoded transposase is to specifically bind to its cognate TIRs that flank at each end of the transposon. This specificity ensures that each transposase can only recognize and bind its own TIR (Figure 1 and Figure 7). Also, binding to the TIR orients the catalytic domain for cleavage, which is normally within the outside end of the TIR (Figure 7). Dimerization of the DBD has been shown to pair the transposon ends to form a synaptic complex, or paired-end complex, such as Mos1 transposase (Richardson et al. 2009). The synaptic complex helps to coordinate cleavage at either end of the transposon.

Since ancestral *Hsmar1* entered the primate genome approximately 50 million years ago, the DBD has been under a strong purifying selection. The DBDs of *Hsmar1* and SETMAR are highly conserved, with only two substituted amino acid residues: K330 and N426 in SETMAR, which are equivalent to E2 and D98 in *Hsmar1*, respectively (Robertson and Zumpano 1997, Cordaux et al. 2006, Miskey et al. 2007). In our TIR complex crystal structure, K330 is located on the N-terminus of the protein near the start of the first α helix. N426 is in the loop connecting α5 and α6. Neither of them interacts directly with the TIR DNA. Substitution of these two residues is unlikely to alter the structure significantly. Thus, the structure of the *Hsmar1* TIR bound SETMAR DBD complex provides a basis for analyzing the sequence-specific binding activity of the

ancestral *Hsmar1* transposase (hereafter *Hsmar1* DBD and SETMAR DBD are used interchangeably).

Overall, the HTH2 motif contributes more nucleobase-specific interactions than that of the HTH1 motif (Figure 21). In the HTHT1 motif, only one residue, R43 (R371 in SETMAR), makes nucleobase-specific interaction with TIR, while in HTH2 four residues R89, H99, S100, and R104 (R417, H427, S428, and R432 in SETMAR) interact with nucleobases (Figure 33). Both HTH motifs bind in the major groove of the TIR DNA (Figure 18).

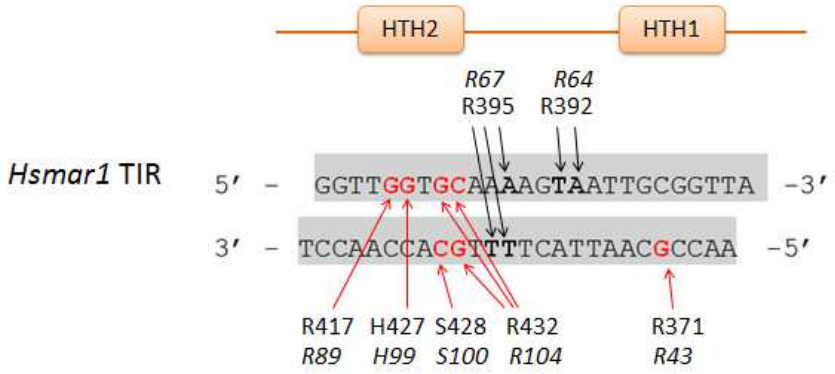

Figure 33. Schematic diagram of sequence-specific binding of *Hsmar1*/SETMAR DNA-binding domain bound to *Hsmar1* TIR.
The simplified HTH1 and HTH2 motifs are presented on the top to show relative positions of the residues. Red arrows indicate the nucleobase-specific interactions. Black arrows indicate minor groove interactions from two AT-hook motifs. The amino acid numberings in the *Hsmar1* transposase are shown as italics.

Our structural results now provide a detailed analysis of specific nucleobase interactions involved in formation of the SETMAR-DNA complex and are overall consistent with the reported 19-bp MBS identified by EMSA (Cordaux et al. 2006) in that all of the nucleobase specific contacts are contained within this site. Important

contributions to the overall binding affinity of SETMAR for either the TIR or variant TIR DNA sequences were assessed by mutation of the nucleobases in direct contact with the protein. The results of these binding assays were consistent with our structural analysis and confirm the assignment of minimal DNA sequence requirements for recognition by SETMAR.

Crystal structures of the DBD or transposase complexed to cognate TIR DNA have been reported for two related transposases (Watkins et al. 2004, Richardson et al. 2009). *Mos1,* an active *mariner* transposon that was first identified in *Drosophila mauritiana* (Jacobson et al. 1986) and *Tc3*, a Tc1 transposon in the *Caenorhabditis elegans* genome (van Luenen et al. 1994). The architecture of the Mos1 DBD is similar to that of *Hsmar1*, with HTH motifs of similar size. However, *the* Mos1 DBD lacks AT-hook motifs found in *Hsmar1* and has a major groove recognition pattern in which two residues of each HTH motif bind in the major groove (Figure 34). Interestingly, the Tc3 DBD includes two small HTH motifs similar to HTH2 in *Hsmar1* and has two AT-hooks (R54 and R57) in the linker region that bind the minor groove of DNA. Unlike *Hsmar1*, however, Tc3 has more nucleobase-specific contact residues in HTH1 than that in HTH2. There are four in HTH1 and two in HTH2 (Watkins et al. 2004) (Figure 34).

The DNA sequences of TIRs are not conserved among Tc1/*mariner* family members (Figure 7); thus, each transposase recognizes the TIR of its own transposon specifically and acts only upon its own transposon ends. Differences in nucleobase-specific contacts

among *Hsmar1*, Mos1 and Tc3 reveal the structural basis of DNA recognition in the Tc1/mariner transposons.

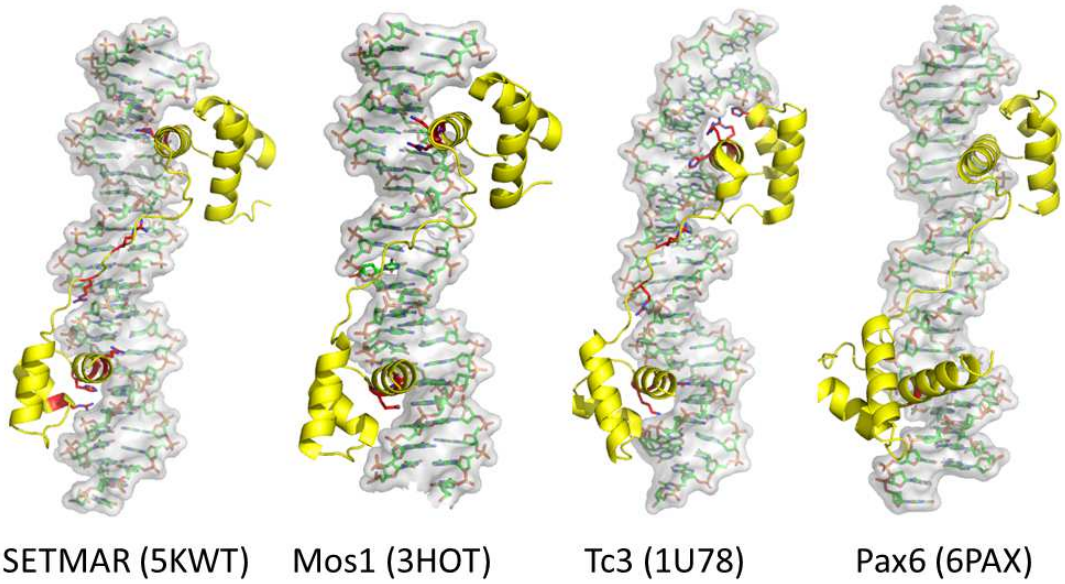

Figure 34. Comparison of related DNA binding domain structures.
The sequence-specific amino acid residues and AT-hook residues are represented as stick models in red color. PDB IDs are in parenthesis. Proteins are rendered as yellow ribbons. DNAs are shown as stick models with surface rendering.

Similar to other members of the Tc1/mariner, *Hsmar1* DBD forms a dimer. The dimeric interface is observed in the HTH1 motif, burying 1610 $Å^2$ of the accessible surface (Figure 19). Hydrophobic residues from helices α1 and α2 constitute the interface. By analogy with the other known structures and studies on Mos1 and Tc3 transposons (Watkins et al. 2004, Richardson et al. 2009), we infer from our structural analysis that the *Hsmar1* transposon functions through a similar mechanism involving a dimeric DBD interaction with TIRs. It is of interest that SETMAR lacks the TIR-specific cleavage activity, examined by different transposition assay systems (Liu et al. 2007,

Miskey et al. 2007) but still retains the dimeric structure that brings the two ends of the transposon together to form a synaptic complex as its ancestors did (Liu et al. 2007).

B. SETMAR binds DNA with a conserved set of residues from *Hsmar1*

In the process of molecular domestication of the *Hsmar1* transposase gene downstream of the *SET* gene, a previously noncoding sequence between these two current exons was converted to encode a linker region. This linker region might serve as an N-terminal extension of the DBD by folding back to interact with the minor groove thereby providing another DNA-interacting element as seen in the Tc3-DNA complex (Watkins et al. 2004) (Figure 34). To explore this possibility in SETMAR, we made a DBD protein including 316-440, 14 amino acids longer than the 329-440 construct. In our structure of this DBD bound to DNA, the N-terminal residues 316-330 are disordered, and thus apparently are not involved in recognition of the DNA.

However, a more critical insight comes from the structural analysis of the variant-TIR complex, obtained from co-crystallization of 316-440 with the variant-TIR DNA. Besides R371 in the HTH1 motif, specific interactions of conserved S428 and R432 involving C and G nucleobases, respectively, are indicated in the variant-TIR complex structure (Figure 24). This conserved set of sequence-specific residues is sufficient to retain high affinity binding to SETMAR in the variant-TIR complex (Figure 27), expanding considerably the number of potential binding sites in the human genome.

C. Biological significance of SETMAR DBD

Transposons were commonly viewed as selfish or molecular parasitic entities to their host (Feschotte 2008, Sinzelle et al. 2009). However, accumulating studies have suggested that transposons have been a rich source of new DNA binding sites. For example, nearly 25% of known human promoters are derived from transposons (Jordan et al. 2003). In the human genome, it has been predicted that there are 1500-7000 SETMAR binding sites (Robertson and Zumpano 1997, Cordaux et al. 2006). This pool of SETMAR binding sites hasn't been systematically explored for their potential biological functions.

Another beneficial effect of transposons is the creation of chimeric genes through molecular domestication (Quesneville et al. 2005, Casola et al. 2008, Sinzelle et al. 2009). The host recruits transposase-derived coding sequences, of which most are the DBD (Breitling and Gerber 2000, Babu et al. 2006, Feschotte 2008, Sinzelle et al. 2009). SETMAR is added to the growing list of transposase-derived DBD proteins. It is proposed that these DBD proteins are recruited as gene regulators (Feschotte 2008). PAX proteins, for instance, share a common ancestor as the members of Tc1/*mariner* superfamily (Breitling and Gerber 2000). They play key regulatory roles in early animal development, and missense mutations within the paired domain have been shown to cause disease (Hanson et al. 1994, Azuma et al. 1996). Interestingly, the DBD comprising two HTH motifs separated by a linker and referred to as a paired domain is found in a large number of protein families, such as Tc1/*mariner* superfamily and PAX family members.

The PAX6 DBD structure (Xu et al. 1999) includes yet another variation of the paired domain in which HTH2 is larger, similar to HTH1 in SETMAR, and HTH1 is smaller like HTH2 in SETMAR (Figure 34). There are no nucleobase-specific residues with HTH1 and only one in the HTH2 motif (Figure 34).

Given that SETMAR binds tightly to TIR DNA *in vitro* (Figure 27), we examined the DNA-binding activity in cell-based assays. In the presence of the TIR binding site, SETMAR represses expression of luciferase in a reporter assay (Figure 30). SETMAR also binds the vector containing TIR sequence as assessed by a ChIP assay (Figure 31), consistent with both the DNA-binding and structural studies.

The variant-TIR complex structure guided selection of a sequence within the promoter region of the *TOPBP1* gene as a potential genomic binding site from cisRED (Robertson et al. 2006). TOPBP1 was of considerable interest as previous studies have shown that overexpression of SETMAR in HEK293T cells promotes restart of stalled replication forks (De Haro et al. 2010). Through ChIP analysis, SETMAR was shown to bind to the promoter region of *TOPBP1* (Figure 4C). Thus, we conclude that SETMAR may play a key role in a number of gene regulatory networks through its function as a transcription factor.

REFERENCES

Adams, P. D., P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger and P. H. Zwart (2010). "PHENIX: a comprehensive Python-based system for macromolecular structure solution." Acta Crystallogr D Biol Crystallogr **66**(Pt 2): 213-221.

Aravind, L. and D. Landsman (1998). "AT-hook motifs identified in a wide variety of DNA-binding proteins." Nucleic Acids Res **26**(19): 4413-4421.

Azuma, N., S. Nishina, H. Yanagisawa, T. Okuyama and M. Yamada (1996). "PAX6 missense mutation in isolated foveal hypoplasia." Nat Genet **13**(2): 141-142.

Babu, M. M., L. M. Iyer, S. Balaji and L. Aravind (2006). "The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons." Nucleic Acids Res **34**(22): 6505-6520.

Bao, W., K. K. Kojima and O. Kohany (2015). "Repbase Update, a database of repetitive elements in eukaryotic genomes." Mob DNA **6**: 11.

Breitling, R. and J. K. Gerber (2000). "Origin of the paired domain." Dev Genes Evol **210**(12): 644-650.

Bricogne G., B. E., Brandl M., Flensburg C., Keller P., Paciorek W., and S. A. Roversi P, Smart O.S., Vonrhein C., Womack T.O. (2016). "BUSTER version 2.10.2 " Cambridge, United Kingdom: Global Phasing Ltd.

Carlson, S. M., K. E. Moore, S. M. Sankaran, J. E. Elias and O. Gozani (2015). "A Proteomic Strategy Identifies Lysine Methylation of Splicing Factor snRNP70 by SETMAR." J Biol Chem.

Casola, C., D. Hucks and C. Feschotte (2008). "Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals." Mol Biol Evol **25**(1): 29-41.

Cordaux, R., S. Udit, M. A. Batzer and C. Feschotte (2006). "Birth of a chimeric primate gene by capture of the transposase gene from a mobile element." Proc Natl Acad Sci U S A **103**(21): 8101-8106.

De Haro, L. P., J. Wray, E. A. Williamson, S. T. Durant, L. Corwin, A. C. Gentry, N. Osheroff, S. H. Lee, R. Hromas and J. A. Nickoloff (2010). "Metnase promotes restart and repair of stalled and collapsed replication forks." Nucleic Acids Res **38**(17): 5681-5691.

Eide, D. and P. Anderson (1985). "Transposition of Tc1 in the nematode Caenorhabditis elegans." Proc Natl Acad Sci U S A **82**(6): 1756-1760.

Emmons, S. W., B. Rosenzweig and D. Hirsh (1980). "Arrangement of repeated sequences in the DNA of the nematode Caenorhabditis elegans." J Mol Biol **144**(4): 481-500.

Emmons, S. W., L. Yesner, K. S. Ruan and D. Katzenberg (1983). "Evidence for a transposon in Caenorhabditis elegans." Cell **32**(1): 55-65.

Emsley, P., B. Lohkamp, W. G. Scott and K. Cowtan (2010). "Features and development of Coot." Acta Crystallogr D Biol Crystallogr **66**(Pt 4): 486-501.

Feschotte, C. (2008). "Transposable elements and the evolution of regulatory networks." Nat Rev Genet **9**(5): 397-405.

Foadi, J., P. Aller, Y. Alguel, A. Cameron, D. Axford, R. L. Owen, W. Armour, D. G. Waterman, S. Iwata and G. Evans (2013). "Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography." Acta Crystallogr D Biol Crystallogr **69**(Pt 8): 1617-1632.

Goodwin, K. D., H. He, T. Imasaki, S. H. Lee and M. M. Georgiadis (2010). "Crystal structure of the human Hsmar1-derived transposase domain in the DNA repair enzyme Metnase." Biochemistry **49**(27): 5705-5713.

Hanson, I. M., J. M. Fletcher, T. Jordan, A. Brown, D. Taylor, R. J. Adams, H. H. Punnett and V. van Heyningen (1994). "Mutations at the PAX6 locus are found in heterogeneous anterior segment malformations including Peters' anomaly." Nat Genet **6**(2): 168-173.

Higgins, J. J., J. Pucilowska, R. Q. Lombardi and J. P. Rooney (2004). "Candidate genes for recessive non-syndromic mental retardation on chromosome 3p (MRT2A)." Clin Genet **65**(6): 496-500.

Ivics, Z., P. B. Hackett, R. H. Plasterk and Z. Izsvak (1997). "Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells." Cell **91**(4): 501-510.

Jacobson, J. W. (1990). "Mariner, Mos and associated aberrant traits in Drosophila mauritiana." Genet Res **55**(3): 153-158.

Jacobson, J. W., M. M. Medhora and D. L. Hartl (1986). "Molecular structure of a somatically unstable transposable element in Drosophila." Proc Natl Acad Sci U S A **83**(22): 8684-8688.

Jeyaratnam, D. C., B. S. Baduin, M. C. Hansen, M. Hansen, J. M. Jorgensen, A. Aggerholm, H. B. Ommen, P. Hokland and C. G. Nyvold (2014). "Delineation of known and new transcript variants of the SETMAR (Metnase) gene and the expression profile in hematologic neoplasms." Exp Hematol **42**(6): 448-456 e444.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany and J. Walichiewicz (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenet Genome Res **110**(1-4): 462-467.

Kim, H. S., Q. Chen, S. K. Kim, J. A. Nickoloff, R. Hromas, M. M. Georgiadis and S. H. Lee (2014). "The DDN Catalytic Motif Is Required for Metnase Functions in Non-homologous End Joining (NHEJ) Repair and Replication Restart." J Biol Chem **289**(15): 10930-10938.

Lampe, D. J., B. J. Akerley, E. J. Rubin, J. J. Mekalanos and H. M. Robertson (1999). "Hyperactive transposase mutants of the Himar1 mariner transposon." Proc Natl Acad Sci U S A **96**(20): 11428-11433.

Lampe, D. J., M. E. Churchill and H. M. Robertson (1996). "A purified mariner transposase is sufficient to mediate transposition in vitro." EMBO J **15**(19): 5470-5479.

Lander, E. S., et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lee, S. H., M. Oshige, S. T. Durant, K. K. Rasila, E. A. Williamson, H. Ramsey, L. Kwan, J. A. Nickoloff and R. Hromas (2005). "The SET domain protein Metnase mediates foreign DNA integration and links integration to nonhomologous end-joining repair." Proc Natl Acad Sci U S A **102**(50): 18075-18080.

Liu, D., J. Bischerour, A. Siddique, N. Buisine, Y. Bigot and R. Chalmers (2007). "The human SETMAR protein preserves most of the activities of the ancestral Hsmar1 transposase." Mol Cell Biol **27**(3): 1125-1132.

Lohe, A. R., D. De Aguiar and D. L. Hartl (1997). "Mutations in the mariner transposase: the D,D(35)E consensus sequence is nonfunctional." Proc Natl Acad Sci U S A **94**(4): 1293-1297.

McCoy, A. J., R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni and R. J. Read (2007). "Phaser crystallographic software." J Appl Crystallogr **40**(Pt 4): 658-674.

Minor, W., M. Cymborowski, Z. Otwinowski and M. Chruszcz (2006). "HKL-3000: the integration of data reduction and structure solution--from diffraction images to an initial model in minutes." Acta Crystallogr D Biol Crystallogr **62**(Pt 8): 859-866.

Miskey, C., Z. Izsvak, R. H. Plasterk and Z. Ivics (2003). "The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells." Nucleic Acids Res **31**(23): 6873-6881.

Miskey, C., B. Papp, L. Mates, L. Sinzelle, H. Keller, Z. Izsvak and Z. Ivics (2007). "The ancient mariner sails again: transposition of the human Hsmar1 element by a

reconstructed transposase and activities of the SETMAR protein on transposon ends." Mol Cell Biol **27**(12): 4589-4600.

Montano, S. P. and P. A. Rice (2011). "Moving DNA around: DNA transposition and retroviral integration." Curr Opin Struct Biol **21**(3): 370-378.

Morgan, G. T. (1995). "Identification in the human genome of mobile elements spread by DNA-mediated transposition." J Mol Biol **254**(1): 1-5.

Mossessova, E. and C. D. Lima (2000). "Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast." Mol Cell **5**(5): 865-876.

Murshudov, G. N., A. A. Vagin and E. J. Dodson (1997). "Refinement of macromolecular structures by the maximum-likelihood method." Acta Crystallogr D Biol Crystallogr **53**(Pt 3): 240-255.

Nowotny, M. (2009). "Retroviral integrase superfamily: the structural perspective." EMBO Rep **10**(2): 144-151.

Oosumi, T., W. R. Belknap and B. Garlick (1995). "Mariner transposons in humans." Nature **378**(6558): 672.

Pace, J. K., 2nd and C. Feschotte (2007). "The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage." Genome Res **17**(4): 422-432.

Painter, J. and E. A. Merritt (2006). "Optimal description of a protein structure in terms of multiple groups undergoing TLS motion." Acta Crystallogr D Biol Crystallogr **62**(Pt 4): 439-450.

Painter, J. and E. A. Merritt (2006). "TLSMD web server for the generation of multi-group TLS models." Journal of Applied Crystallography **39**: 109-111.

Plasterk, R. H. A. (1996). "The Tc1/mariner transposon family." Curr. Top. Microbiol. Immunol. **204**: 125-143.

Quesneville, H., D. Nouaud and D. Anxolabehere (2005). "Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element." Mol Biol Evol **22**(3): 741-746.

Richardson, J. M., S. D. Colloms, D. J. Finnegan and M. D. Walkinshaw (2009). "Molecular Architecture of the Mos1 Paired-End Complex: The Structural Basis of DNA Transposition in a Eukaryote." Cell **138**(6): 1096-1108.

Robertson, G., M. Bilenky, K. Lin, A. He, W. Yuen, M. Dagpinar, R. Varhol, K. Teague, O. L. Griffith, X. Zhang, Y. Pan, M. Hassel, M. C. Sleumer, W. Pan, E. D. Pleasance, M. Chuang, H. Hao, Y. Y. Li, N. Robertson, C. Fjell, B. Li, S. B. Montgomery, T. Astakhova, J. Zhou, J. Sander, A. S. Siddiqui and S. J. Jones (2006). "cisRED: a database system for genome-scale computational discovery of regulatory elements." Nucleic Acids Res **34**(Database issue): D68-73.

Robertson, H. M. (1993). "The mariner transposable element is widespread in insects." Nature **362**(6417): 241-245.

Robertson, H. M. (1995). "The Tc1-Mariner Superfamily of Transposons in Animals." Journal of Insect Physiology **41**(2): 99-105.

Robertson, H. M. and D. J. Lampe (1995). "Distribution of transposable elements in arthropods." Annu Rev Entomol **40**: 333-357.

Robertson, H. M. and R. Martos (1997). "Molecular evolution of the second ancient human mariner transposon, Hsmar2, illustrates patterns of neutral evolution in the human genome lineage." Gene **205**(1-2): 219-228.

Robertson, H. M. and K. L. Zumpano (1997). "Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome." Gene **205**(1-2): 203-217.

Roman, Y., M. Oshige, Y. J. Lee, K. Goodwin, M. M. Georgiadis, R. A. Hromas and S. H. Lee (2007). "Biochemical characterization of a SET and transposase fusion protein, Metnase: its DNA binding and DNA cleavage activity." Biochemistry **46**(40): 11369-11376.

Sinzelle, L., Z. Izsvak and Z. Ivics (2009). "Molecular domestication of transposable elements: from detrimental parasites to useful host genes." Cell Mol Life Sci **66**(6): 1073-1093.

Smit, A. F. (1999). "Interspersed repeats and other mementos of transposable elements in mammalian genomes." Curr Opin Genet Dev **9**(6): 657-663.

Smit, A. F. and A. D. Riggs (1996). "Tiggers and other DNA transposon fossils in the human genome." Proc Natl Acad Sci U S A **93**(4): 1443-1448.

Teske, B. F., M. E. Fusakio, D. Zhou, J. Shan, J. N. McClintick, M. S. Kilberg and R. C. Wek (2013). "CHOP induces activating transcription factor 5 (ATF5) to trigger apoptosis in response to perturbations in protein homeostasis." Mol Biol Cell **24**(15): 2477-2490.

Van Duyne, G. D., R. F. Standaert, P. A. Karplus, S. L. Schreiber and J. Clardy (1993). "Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin." J Mol Biol **229**(1): 105-124.

van Luenen, H. G., S. D. Colloms and R. H. Plasterk (1994). "The mechanism of transposition of Tc3 in C. elegans." Cell **79**(2): 293-301.

Watkins, S., G. van Pouderoyen and T. K. Sixma (2004). "Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA." Nucleic Acids Res **32**(14): 4306-4312.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel and A. H. Schulman (2007). "A unified classification system for eukaryotic transposable elements." Nat Rev Genet **8**(12): 973-982.

Williamson, E. A., K. K. Rasila, L. K. Corwin, J. Wray, B. D. Beck, V. Severns, C. Mobarak, S. H. Lee, J. A. Nickoloff and R. Hromas (2008). "The SET and transposase domain protein Metnase enhances chromosome decatenation: regulation by automethylation." Nucleic Acids Res **36**(18): 5822-5831.

Winn, M. D., C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S.

Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin and K. S. Wilson (2011). "Overview of the CCP4 suite and current developments." Acta Crystallogr D Biol Crystallogr **67**(Pt 4): 235-242.

Wray, J., E. A. Williamson, S. Chester, J. Farrington, R. Sterk, D. M. Weinstock, M. Jasin, S. H. Lee, J. A. Nickoloff and R. Hromas (2010). "The transposase domain protein Metnase/SETMAR suppresses chromosomal translocations." Cancer Genet Cytogenet **200**(2): 184-190.

Wray, J., E. A. Williamson, M. Royce, M. Shaheen, B. D. Beck, S. H. Lee, J. A. Nickoloff and R. Hromas (2009). "Metnase mediates resistance to topoisomerase II inhibitors in breast cancer cells." PLoS One **4**(4): e5323.

Wray, J., E. A. Williamson, S. Sheema, S. H. Lee, E. Libby, C. L. Willman, J. A. Nickoloff and R. Hromas (2009). "Metnase mediates chromosome decatenation in acute leukemia cells." Blood **114**(9): 1852-1858.

Xu, H. E., M. A. Rould, W. Xu, J. A. Epstein, R. L. Maas and C. O. Pabo (1999). "Crystal structure of the human Pax6 paired domain-DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding." Genes Dev **13**(10): 1263-1275.

CURRICULUM VITAE

Qiujia Chen

EDUCATION AND TRAINING:

2010---2016 Ph. D. in Biochemistry and Molecular Biology, Indiana University, Indianapolis, IN
Advisor: Dr. Millie M. Georgiadis
Committee: Drs Thomas D. Hurley, Ronald C. Wek, John J. Turchi, and Mark R. Kelley

2008 X-ray Methods in Structural Biology, CSHL course, Institute of Biophysics (IBP), Chinese Academy of Sciences, Beijing, China

2002---2006 B.S. in Biotechnology, Shantou University, Shantou, Guangdong, China

RESEARCH EXPERIENCE:

2011---2016 **Research Assistant**
Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN
Advisor: Dr. Millie M. Georgiadis

- Structural analysis of SETMAR DNA binding domain and its biological functions.
- Crystallographic and enzymological studies of human apurinic/apyrimidinic endonuclease 1 (hAPE1).

2006—2010 **Research Assistant**
Structural Biology Laboratory, Guangzhou Institute of Biomedicine and Health, Chinese Academy of Sciences (GIBH-CAS), Guangzhou, China
Advisor: Dr. Jinsong Liu

- Recombinant protein expression and purification for structural studies
- Inclusion body purification and refolding

PUBLICATIONS:

2016 **Qiujia Chen**, Michael E. Fusakio, Suk-Hee Lee, Ronald C. Wek, and Millie M. Georgiadis. (2016) A role for transposase-derived SETMAR in gene regulation: insights from crystal structures of DNA-bound Complexes. (submitted to Nature Structural and Molecular Biology)

Qiujia Chen, Millie M. Georgiadis. (2016) Crystallization and phasing of SETMAR-DNA complexes. (submitted to Acta Crystallogr F Struct Biol Commun)

Millie M. Georgiadis, **Qiujia Chen**, Randall Wireman, Jingwei Meng, Chunlu Guo, April Reed, Michael R. Vasko, and Mark R. Kelley. (2016) Small molecule activation of apurinic/apyrimidinic endonuclease 1 reduces DNA damage induced by cisplatin in cultured sensory neurons. DNA Repair 41: 32-41

2014 Hongzhen He #, **Qiujia Chen #** and Millie M. Georgiadis. (2014) High-resolution crystal structures reveal plasticity in the metal binding site of apurinic/apyrimidinic endonuclease I. Biochemistry 53(41): 6520-6529. # co-first author

Hyun-Suk Kim, **Qiujia Chen**, Sung-Kyung Kim, Jac A. Nickoloff, Robert Hromas, Millie M. Georgiadis, and Suk-Hee Lee. (2014) The DDN Catalytic Motif Is Required for Metnase Functions in Non-homologous End Joining (NHEJ) Repair and Replication Restart. J Biol Chem 289(15): 10930-10938.

2013 Jun Zhang, Meihua Luo, Daniela Marasco, Derek Logsdon, Kaice LaFavers, **Qiujia Chen**, April Reed, Mark R. Kelley, Michael L. Gross, and Millie M. Georgiadis. (2013) Inhibition of apurinic/apyrimidinic endonuclease I's redox activity revisited. Biochemistry 52(17): 2955-66.

2012 Meihua Luo, Jun Zhang, Hongzhen He, Dian Su, **Qiujia Chen**, Michael L Gross, Mark R Kelley, Millie M Georgiadis. (2012) Characterization of the redox activity and disulfide bond formation in apurinic/apyrimidinic endonuclease. Biochemistry 51: 695-705.

2011 Yujie Zhang, Tingting Xu, **Qiujia Chen**, Bing Wang, Jinsong Liu (2011) Expression, purification, and refolding of active human and mouse secreted group IIE phospholipase A2. Protein Expression and Purification 80: 68-73

ABSTRACTS:

2016 66th American Crystallographic Association (ACA) Annual Meeting, Denver, CO

Title: A role for SETMAR in gene regulation: insights from crystal structures of the DNA-binding domain in complex with DNA

Author: Qiujia Chen, Millie M. Georgiadis

Hitchhiker's Guide to Biomolecular Galaxy, Purdue University, West Lafayette, IN

Title: Potential role of SETMAR in gene regulation: insights from structural analysis of the DNA-binding domain in complex with DNA

Author: Qiujia Chen, Millie M. Georgiadis

Statewide Structural Biology Forum, Indiana Clinical and Translational Sciences Institute (CTSI), Indianapolis, IN

Title: Structural Basis for DNA Recognition by the DNA Repair Protein SETMAR

Author: Qiujia Chen, Suk-Hee Lee, and Millie M. Georgiadis

2015 17th Annual Midwest DNA Repair Symposium, Indiana University, Bloomington, IN

Title: Structural Basis for DNA Recognition by the DNA Repair Protein SETMAR

Author: Qiujia Chen, Suk-Hee Lee, and Millie M. Georgiadis

Cancer Research Day, Indiana University Simon Cancer Center, Indianapolis, IN

Title: Crystal Structures of the SETMAR DNA-binding Domain Bound to DNA

Author: Qiujia Chen, Suk-Hee Lee, and Millie M. Georgiadis

2014 23rd Congress and General Assembly of the International Union of Crystallography (IUCr2014), Montreal, Canada

Title: Preliminary crystallographic analysis of SETMAR bound to DNA

Author: Qiujia Chen, Suk-Hee Lee, and Millie M. Georgiadis

ORAL PRESENTATION:

2016 66th American Crystallographic Association (ACA) Annual Meeting, Denver, CO

Title: A role for SETMAR in gene regulation: insights from crystal structures of the DNA-binding domain in complex with DNA

Author: Qiujia Chen, Millie M. Georgiadis

2015 17th Annual Midwest DNA Repair Symposium, Indiana University, Bloomington, IN

Title: "Structural Basis for DNA Recognition by the DNA Repair Protein SETMAR"

Author: Qiujia Chen, Suk-Hee Lee, and Millie M. Georgiadis

TRAVEL GRANTS:

2016 IUSM Graduate Student Travel Grant to Attend the 66th American Crystallographic Association (ACA) Annual Meeting ($500) COMPETITIVE
Indiana University School of Medicine, Indianapolis, IN

Travel Grant to Attend the Hitchhiker's Guide to Biomolecular Galaxy Symposium ($50)
Purdue University, West Lafayette, IN

TEACHING EXPERIENCE:

2014---2015 Mentored an undergraduate student in a 6-month capstone research project

2014 Supervised an undergraduate student in an 8-week research project

2014 Tutored high school students in summer

2011, 2012 Trained first-year graduate students during lab rotations